On the dissociation of structural and linear operations in sentence processing

Christos-Nikolaos Zacharopoulos

On the dissociation of structural and linear operations in sentence processing.

by

Christos-Nikolaos Zacharopoulos

University of Sorbonne

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

March 9, 2022

Keywords: language, brain, subject-verb agreement, m/eeg, neural networks.

In accordance with the requirements of the degree of Doctor of Philosophy in the University of Sorbonne, I present the following thesis entitled,

On the dissociation of structural and linear operations in sentence processing.

This work was performed under the supervision of Prof. Stanislas Dehaene. I declare that the work submitted in this thesis is my own, except as acknowledged in the text and footnotes, and has not been previously submitted for a degree at University of Sorbonne or any other institution.

Christos-Nikolaos Zacharopoulos

In memory of grandpa Apostolos and grandma Chrysa.

Acknowledgements

First and foremost, I would like to thank my advisor, Stanislas Dehaene. I was truly lucky to be able to work, and learn, alongside a giant of cognitive neuroscience, and a brilliant mind of such calibre. I would also like to give my sincere gratitude to Yair Lakretz, who was by my side since the beginning. Yair is a researcher with an amazing breath and depth of knowledge in many fields, from machine learning to cognitive theories, and I learned a lot by his side. I am sure that in a few years from now, he will be leading the research front on the intersection of cognitive neuroscience and natural language processing.

Within the lab, I was lucky to be surrounded by a group of wonderful people. Antonio, thanks for being my "assurance manager", without your help, I wouldn't be writing these lines. Vanna, thanks for standing by my side since day one. You are making the lab a better place, and we are lucky to have you. Fosca, thanks for being present in good and bad times. I will always remember the evening working sessions at your place. Fanis, thanks for introducing me to the fascinating world of hippocampal operations, and I hope that one day this project pays off. Until then, see you in the tennis court! Mathias & Théo, I had so much fun working on our project and this was, definitely, the highlight of my PhD. I deeply believe in both of you, and I was lucky to meet you. Christophe & Minye, thanks for always having an open door, and answering all of my questions. Lastly, I would like to thank the nurses of NeuroSpin, and in particular, Véronique, Gaëlle and Laurence.

But, in this journey, I had companions outside the lab. Nial, I cannot even begin to express my gratitude for all the things you did. You were the pilot in all of my experiments, you circulated my fliers, you brought your friends to participate as subjects, and we recorded auditory stimuli together. From the bottom of my heart, thank you. We became good friends through this, and this might be one of the best outcomes of my PhD. Ailsa, Lindsey, thanks for enduring my experiments and for bugging your friends and family to participate. I promise, it was all worth it.

This PhD wouldn't have been possible without the continuous support of my friends and family, to whom I devote this thesis to. First and foremost, I would like to thank my parents Ioannis and Evangelia, and my siblings, Zoe, Apostolos, and Andreas. Lastly, I would like to thank my grandma Maria, who supported my studies since day one. Additionally, I would like to thank my friends Nikoleta, Gregory, Haris, Maria & Hadrien who are now part of our family. Krisztián, Yvonne, Mara, Christina, Teun, I did not forget you! Thanks for being there when it mattered the most.

Last but not least, I would like to thank my partner Revekka. This thesis wouldn't have been written without her support. Revekka, you helped me carry on when all seemed dark, and lifted me up during the most difficult of times. There are simply not enough words to express my gratitude.

Abstract

This thesis seeks to answer an open question in the field of language processing and comprehension: *Is language a phenomenon based on statistical regularities, such as transition probabilities, or is it a phenomenon deeply rooted in a uniquely human ability to represent nested, symbolic structures?* This question essentially contrasts two distinct mechanisms of language processing. The first mechanism is attributed to probabilistic modeling at the sequence level, and presupposes no structural bias (hereafter *linear mechanism*), whereas the second is sensitive to the syntactic structure of the sentence (hereafter *structural mechanism*). Evidence from previous studies on human, non-linguistic sequence processing, and artificial neural language models, suggests that these two mechanisms might co-exist, however, for language processing, their isolation remains a challenge.

To disentangle the two mechanisms, we created a new design that utilizes the classical psycholinguistic phenomenon of subject-verb agreement, and in particular, the modulation of this agreement in the presence of a noun that does not intervene structurally with the agreement configuration (hereafter *attractor*). To identify neural correlates of these mechanisms, we analyzed data collected from two neuroimaging, and one online-behavioral experiment. Additionally, we employed comparisons between human subjects and deep neural networks, to draw a comparative picture between two systems that are assumed to have a different underlying language apparatus.

In experiment 1, we collected behavioral and neural data (EEG & MEG recordings) from 22 human participants and from artificial neural models, presented with the same set of stimuli. The participants performed a forced-choice, violation-detection task, in an RSVP study with an SOA of $500ms$. The experiment was conducted in English and utilized two different grammatical features: number and animacy. In the models, we show the coexistence of both transition- and structure-sensitive effects and for both features. In humans, we failed to trace neural correlates of a phenomenon stemming purely from transition probabilities of linearly- but not structurally adjacent words (i.e: occurring within the embedding of a prepositional phrase; *The boy near the **girls likes** climbing.*). In contrast, when a deviant bigram transition, co-occurred with a syntactically illicit template (i.e: occurring within the embedding of an object-relative clause; *The boy that the **girls likes** leaves.*), the effect was significantly detectable. Notably, we reported attraction effects at the behavioral level, but only for the feature of grammatical number. Taken together, our results suggest that human language processing is dominated by structure-based computations and is largely robust to transition effects. Additionally, our results point to a difference between language processing in humans and neural models, and to a major difference between how humans process sequences of non-linguistic items and sentences.

In experiment 2, we sought to tackle possible confounds that might have led to the null result of a purely Markov-chain based processing in experiment 1. We hypothesized that three factors might be responsible. First, the morphological complexity of English might not have been sufficiently visible to provoke a large violation-of-transition-probability effect, given that the inflectional difference of the singular and the plural tense is mostly based on a single let-

ter. Second, transition-based effects might be transient and therefore dissolved down to an undetectable level in our original slow SOA settings. Third, in experiment 1, the subjects could have employed task-resolution strategies due to the lack of filler trials. To control for such possibilities, we launched a new M/EEG experiment using French, a language with richer inflectional morphology compared to English, utilizing a carefully selected lexicon. We used several shorter stimulus onset asynchronies ($125, 250, 375, 500ms$). Finally, we included filler trials where violations occurred in different places, compared to those of our canonical conditions. This design utilized only the feature of grammatical number. In agreement with our previous results, we managed to decode the effect attributed to the structural mechanism, but not the transition-based effect. When filtering for the correctness of the responses, we managed to trace effects stemming from the parametric, inflectional covariation of the head noun and the attractor. Importantly, these effects were consistently late, surfacing after the onset of the structural effect. Notably, similar to our previous analysis, we could not decode a pure, violation-of-transition-probability effect between linearly-but, not structurally-adjacent words. Experiment 2 thus replicates the findings and conclusions of experiment 1: language processing is driven by structure-based computations and is robust to transition probabilities between adjacent but non-structurally linked words.

Finally, in experiment 3, we sought to isolate correlates of linear processing, by introducing a parametric manipulation of the linear distance between the attractor and the verb. To that end, we utilized a subject relative clause modifier, with an embedded attractor (i.e: *The boy who likes the* ***girl*** *the most* ***sneezes***). We collected behavioral data from a forced-choice, violation-detection RSVP experiment, and compared the performance of the humans to deep neural networks. Driven by our previous analyses, we employed a design with a higher percentage of filler trials, and we carefully selected participants based on their performance on both the filler trials and canonical conditions, in an attempt to reduce the effects of task-resolution strategies. Our results showed a clear effect of linear distance: there was an interference effect when the attractor was adjacent to the verb, but not when it was more distant. Nevertheless, our analysis showed that this effect stems from a dependency realized between the head noun and the attractor, and not between the attractor and the verb. In other words, when the attractor was adjacent to the target verb, we observed a significant effect of congruency, but not of transition. These results corroborate our previous findings and point to a language processing system that robustly circumvents linear effects stemming from transition probabilities between non-structurally adjacent words.

In conclusion, in a series of experiments, we sought to isolate neural and behavioral correlates of two discrete mechanisms in language processing and comprehension. Across all experiments, we consistently detected correlates of the structural mechanism. We also observed an influence of the non-structurally intervening attractor in the resolution of the subject-verb agreement, but importantly, this effect did not stem from a purely transition-based mechanism. Across experiments, we reported attraction effects and identified two asymmetries that bound this attraction phenomenon: first, participants experience mostly grammatical illusions, that is, misjudging an erroneous sentence as grammatical, but not a grammatical sentence as erroneous (grammatical asymmetry); and second, subjects make more errors in the presence of a plural attractor, compared to a singular one.

Our results are in agreement with the cue-based retrieval model of attraction [Wagers et al., 2009], according to which, errors in subject-verb agreement are due to a faulty working-memory mechanism. This model postulates that the parser might retrieve the feature of another element that shares similar morphosyntactic characteristics (attractor), instead of the noun that controls the grammatical agreement (head noun). Our analysis showed that this retrieval mechanism, is indeed, only engaged when the antecedent carries morphosyntactic information, and not

semantic marking, such as animacy. Overall, our results draw a clear picture of a structure-based language processing system, adherent to the recursive nature of syntactic computations postulated by formal syntactic theory.

Abstract in French

Cette thèse cherche à répondre une question ouverte dans le domaine du traitement et de la compréhension du langage: *le langage est-il un phénomène basé sur des régularités statistiques, telles que les probabilités de transition, ou est-ce un phénomène enraciné dans une capacité humaine unique à représenter des structures imbriquées et symboliques?* Cette question met essentiellement en contraste deux mécanismes distincts du traitement du langage. Le premier mécanisme est attribué à la modélisation probabiliste au niveau de la séquence et il ne présuppose aucun biais structurel (ci-après mécanisme *linéaire*), tandis que le deuxième est sensible à la structure syntaxique de la phrase (ci-après mécanisme *structurel*). Les preuves d'études antérieures sur le traitement des séquences humaines non linguistiques et des modèles de langage neuronal artificiel suggèrent que ces deux mécanismes pourraient coexister, mais pour le traitement du langage, leur isolement reste un défi. Pour démêler les deux mécanismes, nous avons créé un nouveau design qui utilise le phénomène psycholinguistique classique de l'accord sujet-verbe, et en particulier, la modulation de cet accord en présence d'un nom qui n'intervient pas structurellement avec la configuration de l'accord (ci-après *attracteur*). Pour identifier les corrélats neuronaux de ces mécanismes, nous avons analysé les données recueillies à partir de deux expériences de neuro-imagerie et d'une expérience comportementale en ligne. De plus, nous avons utilisé des comparaisons entre des sujets humains et des réseaux de neurones profonds pour construire un tableau comparatif entre deux systèmes supposés avoir un appareil linguistique sous-jacent différent.

Dans l'expérience 1, nous avons collecté des données comportementales et neuronales (enregistrements EEG et MEG) de 22 participants humains et de modèles neuronaux artificiels, présentés avec le même ensemble de stimuli. Les participants ont effectué une tâche de détection de violation à choix forcé, dans une étude RSVP avec un SOA de 500 ms. L'expérience a été menée en anglais et a utilisé deux caractéristiques grammaticales différentes: le nombre et l'animéité. Dans les modèles, nous montrons la coexistence des effets sensibles à la transition et à la structure et pour les deux caractéristiques. Chez les humains, nous n'avons pas réussi à tracer les corrélats neuronaux d'un phénomène provenant uniquement des probabilités de transition de mots linéairement adjacents mais pas structurellement(i.e: se produisant dans l'incorporation d'une phrase prépositionnelle; The boy near the **girls likes** climbing.). En revanche, lorsqu'une transition bigramme déviante coexistait avec un modèle syntaxiquement illicite (i.e: se produisant dans l'incorporation d'une clause relative à l'objet ; The boy that the **girls likes** leaves), l'effet était significativement détectable. Nous avons notamment rapporté des effets d'attraction au niveau comportemental, mais uniquement pour la caractéristique du nombre grammatical. Pris ensemble, nos résultats suggèrent que le traitement du langage humain est dominé par des calculs basés sur la structure et il est largement robuste aux effets de transition. De plus, nos résultats indiquent une différence entre le traitement du langage chez les humains et les modèles neuronaux, et une différence majeure entre la façon dont les humains traitent les séquences d'éléments et de phrases non linguistiques.

Dans l'expérience 2, nous avons cherché à aborder les confusions possibles qui auraient pu

conduire au résultat nul d'un traitement exclusivement basé sur la chaîne de Markov dans l'expérience 1. Nous avons émis l'hypothèse que trois facteurs pourraient être responsables. Tout d'abord, il est possible que la complexité morphologique de l'anglais ne soit pas suffisamment visible pour provoquer un important effet de violation de la probabilité de transition, étant donné que la différence flexionnelle du singulier et du pluriel est principalement basée sur une seule lettre. En deuxième lieu, les effets basés sur la transition peuvent être transitoires et donc dissous à un niveau indétectable dans nos paramètres SOA lents d'origine. Troisièmement, dans l'expérience 1, les sujets auraient pu utiliser des stratégies de résolution de tâches en raison du manque d'essais de remplissage. Pour contrôler ces possibilités, nous avons lancé une nouvelle expérience M/EEG utilisant le français, une langue avec une morphologie flexionnelle, plus riche que l'anglais, en utilisant un lexique soigneusement sélectionné. Nous avons utilisé plusieurs asynchrones de début de stimulus plus courts ($125, 250, 375$ & $500ms$). Enfin, nous avons inclus des essais de remplissage où les violations se sont produites à des endroits différents, par rapport à ceux de nos conditions canoniques. Cette conception n'utilisait que la caractéristique du nombre grammatical. En accord avec nos résultats précédents, nous avons réussi à décoder l'effet attribué au mécanisme structurel, mais pas l'effet transitionnel. Lors du filtrage de l'exactitude des réponses, nous avons réussi à retracer les effets découlant de la covariation paramétrique et flexionnelle du nom principal et de l'attracteur. Il est important de noter que ces effets étaient systématiquement tardifs, apparaissant après le début de l'effet structurel. Notamment, comme dans notre analyse précédente, nous n'avons pas pu décoder un pur effet de violation de la probabilité de transition entre des mots linéairement mais pas structurellement adjacents. L'expérience 2 reproduit ainsi les résultats et les conclusions de l'expérience 1 : le traitement du langage est piloté par des calculs basés sur la structure et robuste aux probabilités de transition entre des mots adjacents mais non structurellement liés. Enfin, dans l'expérience 3, nous avons cherché à isoler des corrélats de traitement linéaire, en introduisant une manipulation paramétrique de la distance linéaire entre l'attracteur et le verbe. Dans ce but, nous avons utilisé un modificateur de clause relative du sujet, avec un attracteur intégré (I.e: The boy who likes the girl the most sneezes). Nous avons recueilli des données comportementales à partir d'une expérience RSVP à choix forcé et à détection de violation, et nous avons comparé les performances des humains aux réseaux de neurones profonds. Motivés par nos analyses précédentes, nous avons utilisé une conception avec un pourcentage plus élevé d'essais de remplissage, et nous avons soigneusement sélectionné les participants en fonction de leurs performances à la fois dans les essais de remplissage et dans les conditions canoniques, dans le but de réduire les effets des stratégies de résolution de tâches. Nos résultats ont montré un effet clair de la distance linéaire à la fois sur les temps de réaction et sur le taux d'erreur : il y avait un effet d'interférence lorsque l'attracteur était adjacent au verbe, mais pas lorsqu'il était plus éloigné. Néanmoins, notre analyse a montré que cet effet provient d'une dépendance réalisée entre le nom principal et l'attracteur, et non entre l'attracteur et le verbe. En d'autres termes, lorsque l'attracteur était adjacent au verbe cible, nous avons observé un effet significatif de congruence, mais pas de transition. Ces résultats confirment nos découvertes précédentes et indiquent un système de traitement du langage qui contourne de manière robuste les effets linéaires résultant des probabilités de transition entre des mots non structurellement adjacents.

En conclusion, dans une série d'expériences, nous avons cherché à isoler les corrélats neuronaux et comportementaux de deux mécanismes discrets dans le traitement et la compréhension du langage. Dans toutes les expériences, nous avons systématiquement détecté des corrélats du mécanisme structurel. Nous avons également observé une influence de l'attracteur intervenant non structurellement dans la résolution de l'accord sujet-verbe, mais surtout, cet effet ne découlait pas d'un mécanisme uniquement basé sur la transition. À travers les expériences, nous avons signalé des effets d'attraction et identifié deux asymétries qui lient ce phénomène

d'attraction : premièrement, les participants éprouvent principalement des illusions grammaticales, c'est-à-dire qu'ils jugent à tort une phrase erronée comme grammaticale, mais pas une phrase grammaticale comme erronée (asymétrie grammaticale) ; et deuxièmement, les sujets font plus d'erreurs en présence d'un attracteur pluriel, par rapport à un attracteur singulier. Nos résultats sont en accord avec le modèle d'attraction basé sur les indices, selon lequel les erreurs d'accord sujet-verbe sont dues à un mécanisme de mémoire de travail défectueux. Ce modèle postule que l'analyseur pourrait récupérer la caractéristique d'un autre élément qui partage des caractéristiques morphosyntaxiques similaires (attracteur), au lieu du nom qui contrôle l'accord grammatical (nom principal). Notre analyse a montré que ce mécanisme de récupération n'est en effet engagé que lorsque l'antécédent porte une information morphosyntaxique, et non un marquage sémantique, comme l'animéité. En général, nos résultats dessinent une image claire d'un système de traitement du langage basé sur la structure, adhérent à la nature récursive des calculs syntaxiques postulés par la théorie syntaxique formelle.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 On the nature of neurolinguistic operations.

Language is such an amazing phenomenon, inherently complex and intrinsically beautiful. Yet, many aspects of this uniquely human mechanism remain elusive. In the words of Robbins Burling [Burling, 2007],

> *We know more about the ways the vocal tract produces the sounds of speech than about how the brain deals with them.*

Language appears trivial, but it is nothing but that. Ray Jackendoff in his book *Patterns in the Mind: Language and Human Nature* [Jackendoff, 2008] describes this dichotomy vividly:

> *The main thing is to appreciate how hard a problem this is. The fact that we can talk (and cat's can't) seems so obvious that it hardly bears mention. But just because it's obvious doesn't mean it's easy to explain.*

Language is a matter of contention. Is it an innate ability? Does it purely serve communication? Is it a syntax-dominated mechanism, or is it semantics, the driving force of language? This very controversial nature of language comes as a byproduct of the underlying complexity that language encloses. Of the many unresolved questions within the domain of language studies, a topic that often stirs up heated controversy entails the nature of linguistic computations [Ding et al., 2017, Haskell and MacDonald, 2005, Ding et al., 2015, Willer Gold et al., 2017, Arana et al., 2021]. Is language enabled via an innate, human specific, ability to mentally represent linear input (such as words in a sentence) into structured representations? According to formal linguistic accounts, that is indeed the case. On the other hand, advances on the deep learning front challenge this point of view. Artificial natural language processing models are getting progressively better at resolving linguistic tasks. These models presuppose no structural bias, and operate solely based on statistical learning and probabilistic modeling at the sequence level. Simply put, these models operate on a linear, rather than a structural-basis. This thesis will touch upon this debate, utilizing an experimental design that allows for a direct contrast between structure-based and low-level, linear-order computations, and will seek to answer the following question: *Is language a phenomenon based on statistical regularities, such as transition probabilities, or is it a phenomenon deeply rooted in a uniquely human ability to represent nested, symbolic structures?*

1.1.1 Hierarchical processing

The question regarding the nature of linguistic computations is paramount. The root of human singularity is often attributed to language, and many linguists following the seminal work of

Chomsky believe that language is enabled via the mechanism of recursion[Chomsky, 2014a, Chomsky, 2009, Hauser et al., 2002]. Recursion refers to the ability to apply a function on its own output.[1]. In language, this property is reflected in the ability of the parser to combine, or *merge* [Chomsky, 1957] syntactic units, and apply the same rule repeatedly.

A simple illustration of the recursiveness that governs language is the ability to extend (theoretically, *ad infinitum*) a sentence using relative clauses,

I saw a cat on the park → I saw a cat that chased a mouse on the park → I saw a cat that chased a mouse that was eating cheese on the park → ...

or coordination,

I saw a dog and a cat → I saw a dog and a cat and a cow ...

Evidence for recursion, as the core of human singularity, first came from an influential behavioral study by Fitch & Hauser [Fitch and Hauser, 2004]. The authors collected data from a familiarization-discrimination paradigm, from both human subjects and tamarin monkeys, in a design that compared two types of artificial grammars. The first grammar, called Finite State Grammar (FSG), is simple and can be described by the formula $(AB)^n$. This formula implies that for every element 'A', an element 'B' will follow. In other words, this language can be fully described via means of transition probabilities. The second grammar, called Phrase Structure Grammar (PSG), contains hierarchical structure and can be compressed to the formula A^nB^n. For this grammar, a simple transition-probability based description cannot suffice. The main characteristic of this grammar is recursiveness, therefore, structure. Indeed, Fitch & Hauser showed that humans could master both types of grammars, whereas the non-human primates could only perform well on the FSG.

In a follow-up fMRI study, Angela Friederici and colleagues, compared the two grammars while human subjects performed a violation-detection task[Friederici et al., 2006a]. The subjects had learned the two grammars prior to the task. For the FSG, the authors reported activity in the Frontal Operculum, but not in brodmann area 44 (Broca's area). In contrast, for the PSG, the Frontal-Operculum was still engaged, but there was also activation in BA44. Importantly, both grammars were equally difficult for the subjects. This study hence concluded that the brain is indeed sensitive to structural hierarchy, and this is processed in BA44.

Nevertheless, these results were based on artificial grammars. Friederici and colleagues then sought to identify whether the same holds for natural language [Friederici et al., 2006b]. The authors varied syntactic hierarchy (embedded vs non-embedded) and dependency length (short vs long), in a study that was conducted in German. The authors confirmed that the main effect of hierarchy was located in BA44.

To further verify the active involvement of BA44 in hierarchical processing, the same group investigated whether the activation in this region increases parametrically with sentence complexity [Makuuchi et al., 2009]. Friederici and colleagues verified, with a region of interest analysis, that indeed the BOLD signal change in BA44 was modulated as a function of sentence complexity.

As already mentioned, Noam Chomsky proposed that a fundamental linguistic operation, underlying recursion, is that of *merge* [Chomsky, 2014b]. According to this, two linguistic units can be combined to create a new one (e.g: [[The][Ship]] → [The Ship]). Therefore, direct evidence of such operation would be in line with a sensitivity of hierarchical operations in the human brain. Evidence from two recent fMRI and one ECOG[2] study indeed confirm that.

[1] An illustrating example of recursion comes from mathematics and the *Fibonnaci sequence*, which is defined as: $f(i) = f(i-1) + f(i-2)$

[2] Electrocorticography

In a seminal work, Pallier et al.[Pallier et al., 2011] presented word-sequences of varying lengths to adult volunteers during fMRI. Their analysis led to an identification of a left-lateralized, syntax specific network that includes the inferior frontal gyrus or "Broca's area" (IFG, Brodmann areas BA44 and 45) and the superior temporal sulcus (STS). In these areas, the activity increased monotonically as a function of the complexity of the phrase structure. Notably, the authors also included a "Jabberwocky" condition. In this condition, the sentences resembled normal sentences, but all the non-function words were replaced with meaningless tokens. This led to the designation of brain areas responsible for core-syntax operations. In particular, the authors reported monotonically increasing activation in both frontal, and posterior STS regions, irrespective of lexico-semantic information. In a 2015 fMRI study, Emiliano Zaccarella and Angela Friederici investigated the difference in activations of the subjects brains, when confronted with word-lists versus syntactic phrases [Zaccarella and Friederici, 2015]. The authors pointed to a systematic activation in BA44 for phrase structures, but not for word-lists. Importantly, this result was evident not only at the aggregated group analysis, but also at the single subject level. In an ECOG study, Nelson et al. [Nelson et al., 2017] analyzed high gamma activity from a rapid-serial-visual (RSVP) presentation experiment and identified sets of electrodes within the above-mentioned network, where the activity increased with each consecutive word, but abruptly diminished once a phrase could be completed. This result was interpreted as an indicant of a neural correlate of the 'merge' mechanism.

The above-mentioned studies report results from English, French, and German, but evidence of active engagement of BA44 in syntactic processing is evidence from other languages including Hebrew and Japanese [Friederici, 2011].

Evidence for sensitivity to hierarchical operations does not only come from brain imaging studies, but also from behavioral experiments. Shi et al. showed that toddlers can effectively understand hierarchical phrase structures, necessary to determine the grammatical configuration of two distinct linguistic structures [Shi et al., 2020]. In a recent study comparing humans and artificial language models, Coopmans et al. [Coopmans et al., 2021] showed that humans only interpret equivocal noun phrases such as "second blue ball" using a hierarchical parsing, in contrast to language models that require explicit hierarchical information during training.

Overall, a plethora of brain and behavioral studies as well as theoretical work [Hauser et al., 2002, Berwick and Weinberg, 1986, Everaert et al., 2015, Jackendoff, 1972, Kratzer and Heim, 1998, Partee, 1975, Pinker, 1998, Lidz et al., 2003, Martin, 2020] point to linguistic operations driven purely by hierarchical structure, blind to the surface or linear representation of the sentences. From this standpoint, in language, "what you see is not what you get"[Everaert et al., 2015].

Indeed, according to this view, a sentence can have two representations. A so-called linear form, and a corresponding tree-representation. The linear representation of a sentence is the familiar form used in writing. For example, the following sentence is presented in a linear (or sequential) form:

(1) The boy near the girl likes climbing.

The structure-driven hypothesis of language representation states that this very same sentence can be represented using a syntactic tree, such as this, shown in figure 1.1.

This representation opens the possibility to define two distinct distance metrics. A structural-distance and a linear-distance. We would return to these definitions often in this thesis, but first, we ought to describe an important element of psycholinguistics: agreement.

In order for sentence (1) to be grammatical, the noun "boy" (hereafter *trigger*) must agree with the verb "likes" (hereafter *target*). The feature of the trigger-target agreement, in this case, is grammatical number.

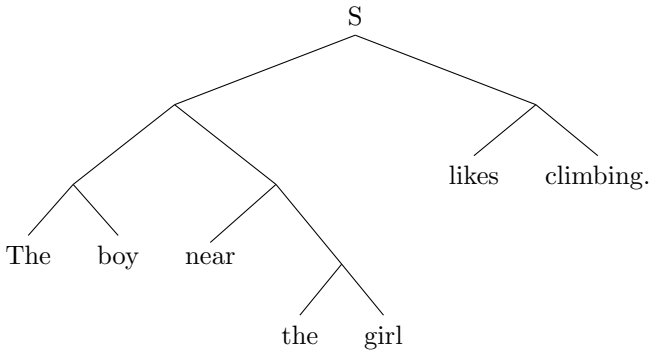


Figure 1.1: The syntactic-tree representation of the sentence.

Hence, in order for the grammaticality of the sentence to be evaluated, the agent (neural or artificial) needs to investigate the agreement relationship between the trigger and the target. According to Chomsky et al. [Berwick et al., 2011], the brain performs that under the principle of *minimal structure distance*. Chomsky [Chomsky, 2015] states that:

> *Language makes use of a property of minimal structural distance, never using the much simpler operation of minimal linear distance; in this and numerous other cases, ease of processing is ignored in the desing of language. In technical terms, the rules are invariably structure-dependent, ignoring linear order*

This argument can be demonstrated easier with an example. Consider the following sentence:

(2) The boy near the girls likes climbing.

In this sentence, there exists a "local" disagreement in the bigram *girls-likes*. Nevertheless, humans are able to judge this sentence as grammatical, despite the interference of a token that disagrees with respect to the feature of interest (grammatical number). This intervening element is called an *attractor*, and the effects investigating these intervening phenomena are called *attraction effects*.

According to Chomsky, the ability to circumvent the attractor interference, is due to the property of the minimal structure distance, which stems from a much broader principle, that of Minimal Computation [Chomsky, 2015]. Based on this approach, the structural trigger-target distance is smaller compared to the distance between the second noun (girls) and the target. In other words, the second noun is deeper into the syntactic tree and thus needs to "transverse" a greater distance to reach the target.

1.1.2 Non-structural processing

The existence of structure as the underlying faculty of language, is often perceived axiomatically [Bybee, 2002, Uddén et al., 2020] and, although broadly, is not universally accepted. In the words of Ioan Bybee [Bybee, 2002]:

> *Linguists rarely ask why natural language has constituent structure; they merely assume that it does, just as they assume that all phonologies will be organized into segments and features.*

The recent success of natural language processing (NLP) and deep learning (DL) is challenging the theory-driven approach sketched above. Most NLP models don't have explicit structural bias and perceive sentences simply as a sequential string of words.

Stefan Frank, Rens Bod and Morten Christiansen in their review paper ¨*How hierarchical is language use?*¨[Frank et al., 2012] make a claim that in terms of computational expenditure, a sequential processing approach should be favored over a hierarchical one, sketching a non-hierarchical model of language processing that relies on sequential order. In line with this view, a growing body of research supports the view that statistical regularities alone could suffice to explain linguistic phenomena usually attributed to hierarchical operations [Elman, 1990, Bybee, 2002, Frank et al., 2012, Frank and Christiansen, 2018b, Christiansen and Chater, 2015, Haskell and MacDonald, 2005, Haskell et al., 2010]. In a behavioral study, Frank and Bod [Frank and Bod, 2011] demonstrated that reading times could be explained by probabilistic modelling alone, thus disregarding the necessity of a hierarchical processing. Operations that do not assume any underlying hierarchy and only occur at the sequence, or word-order level, are often termed as *linear*. The implementations of this mechanism stem from an information-theory approach to language processing, and include metrics such as word-suprisal and entropy. It is important to note, that information-based approaches and structure-based processing are not always mutually exclusive and often assume an underlying phrase-structure grammar [Hale, 2016]. For a review on probabilistic accounts of language processing, see [Armeni et al., 2017]. Therefore, the debate is not whether predictive coding is actively engaged in language processing, but whether prediction occurs at the structural or sequence (linear) level.

1.2 Bridging the gap: Computational modelling and comparisons with human data.

The success of Deep Learning in natural language processing is palpable, with applications ranging from machine translation to speech recognition and text summarization. Nevertheless, prof. Noam Chomsky, believes that this success is simply an engineering feat, irrelevant to science[Norvig, 2017] and the deep questions a scientist should ask about the nature of the world [Chomsky, 2015].

Nevertheless, computational modelling is a valuable tool for cognitive neuroscience and neurolinguistics, as it allows researches to compare the performance of the networks directly with that of humans and generate predictions, but also, and most importantly, researches have the ability to approach the networks "optogenetically", either by ablating certain units or by introducing perturbations to the networks[Lakretz et al., 2019a].

As previously mentioned, number agreement over long-distance, is considered a classical demonstration of structural operations, associated with the recursive ability of the biological system. On what is now considered a hallmark paper, Tal Linzen, Emmanuel Dupoux and Yoav Goldberg evaluated the ability of Recurrent Neural Networks (RNNs) to perform the *number prediction task* [Linzen et al., 2016a] on long-range dependencies. The network was presented with all but the last token, and asked to predict the last work, a process called "language modelling" [Linzen and Baroni, 2021].

Importantly, Linzen and colleagues evaluated the performance of the network in the presence of intervening nouns that carried an opposite grammatical number (attractors). The network achieved high performance (82%) even in sentences with four attractors. The ability of the networks to perform the number agreement task is not constrained to the specific architecture of the RNNS, as many other architectures are equally successful (for a review, see [Linzen and Baroni, 2021]).

On a later study, Lakretz and colleagues [Lakretz et al., 2019a] took a neuroscientific approach to Deep Learning, by opening the "black box" of a Long short-term memory network (LSTM), a specific, gated-architecture of an RNN, and the same kind of architecture that [Linzen et al., 2016a] used.

The authors ablated units of the LSTM (by setting the activation function to zero) and observed the performance of the network in the number agreement task. By doing so, Lakretz and colleagues identified a circuit of neurons in the network that are responsible for performing the task.

There are three types of units in the network that operate in order for the network to correctly conjugate the verb of the sentence over a long distance. The first unit is the so-called Syntax unit. The activity of this unit traces the structure of the long-range dependency. There also exist two types of long-range units, with which the syntax unit has strong synaptic connections. In a dependency where the main noun is singular, the Singular Unit is active throughout, whereas the Plural unit is silent. In contrary, when the main noun of the subject is plural, the Plural unit is active and the Singular is silent. Finally, there exist short-range units that are active throughout but only represent the last encountered number and cannot carry information across the long-range.

A striking aspect of this study is the sparsity of this mechanism. From a total of 1300 units, only a few units are carrying this operation. This experiment was conducted in English, but later replicated in Italian [Lakretz et al., 2021b].

The authors tested many models, different in the initialization of the training procedure. After the ablation of very few units, the networks reach chance level performance on agreement tasks. This not confined to the number feature, but is also confined to different grammatical features, such as gender.

In the same study, the authors moved from a single long-range dependency (The **keys** to the cabinet **are** ...) to two long-range dependencies (The **keys** that the man near the cabinet holds **are** ...) forming a 2×2 factorial design, where the first factor manipulates whether the two dependencies are successive or nested. The second factor manipulates whether the embedded dependency is short or long. Given the sparsity of the mechanism, the authors predicted that the network will not manage to resolve the nested dependency. Additionally, the predictions stated that this should not be the case in the successive dependency, given the ability of the network to process them sequentially.

Indeed, in the successive dependencies, the long-range units process the dependencies sequentially for both the short-range and long-range cases. In the nested cases (both short and long), the units can process robustly the outer dependencies, but not the inner. This was interpreted as evidence for the sparsity of the mechanism, given the inability of the network to allocate computational resources.

The authors then compared the behavioral performance of the network to that of the humans. The network had very low error-rate in the successive dependencies (both short and long). In the nested cases, the network was erroneous, in particular when the main and the embedded noun(s) disagreed in grammatical number.

When comparing the short vs long dependencies, Lakretz and colleagues showed that indeed the networks make more errors on the inner dependency compared to the outer one. This was especially pronounced in the case of long-range dependencies. Interestingly, the performance of the humans followed an extremely similar structure. This might be an indication that humans operate based on a sparse mechanism, similar to the networks. However, and despite the initial similarity of the results, the authors observed one big difference between the human and the network performance. For the models, the performance on the long-nested dependency was below chance level. Therefore, although there is high similarity between humans and the

networks, there is still a fundamental difference in the ability to process complex linguistic constructions.

In yet another recent work by Lakretz et al. [Lakretz et al., 2021a], the authors tested the robustness of syntactic processing in humans and transformed-based language models. Lakretz and colleagues used the same setup as [Lakretz et al., 2021b], and introduced a set of three words to the inner dependency of their design. Surprisingly, the minimal addition brought the transformer's performance to below chance, whereas prior to this addition, these language models achieved "superhuman" performance (almost zero error-rate). The authors interpreted these results as evidence of instability of syntactic processing in state-of-the-art language models, a phenomenon which was not observed in the human subjects.

To sum up, in the results of [Lakretz et al., 2021b], a sparse, long-range mechanism for grammatical agreement consistently emerged in both humans and models, across languages, grammatical features and network random initialization. However, the RNNs fail to process inner dependencies of nested constructions. Humans also make errors in inner dependencies, but perform significantly better compared to RNNs, and above chance level. This difference can be explained by an underlying recursive processing in humans but not in the networks.

Other than behavioral evidence, language modelling can be applied in conjunction with neuroimaging studies. In a 2021 study by Charlotte Caucheteux, Alexandre Gramfort and Jean-Remi King, the authors tested whether the gap between human and artificial behavioral performance can be explained by the range over which the biological and artificial systems are able to make predictions [Caucheteux et al., 2021]. The authors tested the hypothesis that the biological system is optimized to perform long-range and hierarchical predictions, unlike language models, which are fine-tuned to predict adjacent words.

In what can only be considered a "big-data" approach, the authors went on to put this hypothesis to the test, by showing that activations of deep neural network models follow a linear mapping onto the fMRI responses of a total of 345 subjects that listened to short stories.

[Caucheteux et al., 2021] showed that this mapping can be significantly improved if the networks are enhanced with language representations of the next 8 words. Additionally, the mapping of the enhanced model broadly coincides with brain regions associated with language processing. The authors showed a hierarchy of predictions in the human brain, with fronto-parietal regions engaged in long-range predictions, and superior-temporal regions involved in short-term predictions.

Section Summary

The plethora of extremely recent studies in the intersection of neuroscience and computational modelling, is indicative of the benefits of studying both the human and the artificial system. Comparative studies between humans and Deep Neural Networks can efficiently pave the way towards understanding the inner mechanisms of language processing in the human brain.

1.3 Subject-verb agreement and models of attraction

1.3.1 Neural correlates of subject-verb agreement

There exists a long body of research regarding the neural correlates of subject-verb agreement in many languages. Overall, the literature points to an ERP profile that consists of a LAN (Left Anterior Negativity, observed 300-500ms post verb onset) followed by a P600 (Positivity starting at 500ms post verb onset) [Kutas and Hillyard, 1983, Osterhout and Holcomb, 1992, Osterhout and Mobley, 1995, Kaan et al., 2000, Kaan, 2002, Molinaro et al., 2011b, Friederici et al., 1996, Rossi et al., 2005, Hagoort et al., 1993, Hagoort et al., 2003, Friederici, 2017, Molinaro

et al., 2011a, Coulson et al., 1998]. Nevertheless, the LAN reports are irreconcilable across studies. In particular, the LAN has been hypothetized to be a byproduct of the averaging process but also dependent on the complexity of the linguistic construction [Molinaro et al., 2011a] and has been a topic of debate over the past years [Tanner and Van Hell, 2014, Molinaro et al., 2015, Caffarra et al., 2019]. On the other hand, most subject-verb agreement studies report systematically effects in the time-frame of the P600. Another component that has been reported, mostly by the work of Friederici is that of ELAN (Early Left Anterior Negativity emerging between 100 and 300 ms after the onset of the target word) [Friederici and Kotz, 2003, Hahne and Friederici, 1999, Herrmann et al., 2011, Friederici, 2017]. However, this component is also a matter of debate [Steinhauer and Drury, 2012] and seems to be dependent on experimental settings and modality [Molinaro et al., 2011a].

1.3.2 Attraction Phenomena

Agreement attraction occurs when the subject-verb computation is disrupted due to the presence of a distractor noun. Traditionally, this noun is referred to as an *"attractor"*. There is a long-history of attraction studies that can be traced back to the 1960s [Zandvoort, 1961], but most of the influence in the field comes from the works of Kathryn Bock [Bock and Miller, 1991, Bock and Cutting, 1992, Bock and Eberhard, 1993], who provided psycholinguistic evidence for the "Markedness phenomenon" in the case of number agreement.

The Markedness phenomenon

This is a phenomenon that bears polysemy in linguistics. In his influential paper, Martin Haspelmath describes twelve different uses of the term in linguistics [Haspelmath, 2006]. Essentially, this phenomenon refers to the property of certain linguistic elements to be "marked" compared to their binary counterparts. The markedness of these elements can be associated with an increased processing cost.

In the case of number agreement, the marked property is the plural number, and the corresponding unmarked (or default) is the singular. Kathryn Bock [Bock and Miller, 1991] showed that attraction effects occur only when the attractor is plural. This can be better demonstrated with an example, comparing the two sentences,

1. The key to the cabinets
2. The keys to the cabinet

Bock et al. showed that subjects made more errors in a production task, in the first sentence compared to the second. This asymmetry in the behavioral index metrics between the singular and the plural attractor is called a *mismatch asymmetry*. This asymmetry can be thought as an instantiation of the markedness effect for the grammatical number feature. Notably, similar effects have been shown in studies of computational modeling [**?**]

Grammatical Asymmetry

In an influential paper by Matt Wagers, Ellen Lau and Colin Phillips [Wagers et al., 2009], the authors showed the presence of an attractor had a severe effect in the error rate of the agrammatical sentences but a less significant effect in the grammatical cases. In other words, the presence of an attractor in the sentences that contain a violation, leads to an *a-grammatical illusion*[Phillips et al., 2011] but not to *grammatical illusions*. This effect is termed *grammatical asymmetry* and bears great significance for the explanation of the attraction mechanisms. The

explanation of this asymmetry is the crucial differentiating factor between the existing models of attraction.

Models of attraction

There exist two main families of models that account for the attraction phenomena. The main difference between the two, is whether the representation of the head noun of the sentence is active continuously, or whether a memory mechanism is enabled upon a cue, that retrieves the number of the subject from the memory system.

Continuous representation models: In the first family of models, the errors occur because the representation of the subject number is erroneous [Eberhard et al., 2005, Franck et al., 2002, Staub, 2009, Staub, 2010, Vigliocco and Nicol, 1998]. The first account of this line of thinking can be traced back to 1924 and the work of Otto Jespersen [Sonnenschein, 1925], who claimed that:

> *If the verb comes long after the verb, there is no more mental energy to remember what the number of the subject was, and therefore the system uses the number of the closest noun.*

This led Fayol and colleagues, [Fayol et al., 1994] to approach the subject-verb agreement phenomenon through the lens of a spreading of activation. In a seminal 1994 work, the authors claimed that:

> *Agreement is computed automatically through the spreading of activation from the subject to the verb. When there is a local noun, activation will spread from this noun too.*

This model was later formalized into a quantitative version by Ederbard and colleagues under the term: "Marking and Morphing model" [Eberhard et al., 2005]. For a brief review on this model, see [Hammerly et al., 2019]. Another model that assumes a global representation of the subject number is the percolation model. In this account, the number feature of the intervening noun percolates upwards to the root node of the syntactic tree [Franck et al., 2002].

Cue-based memory model: This model is based on a cue-based memory architecture according to which, upon a cue elicited by the verb, the parser would search for the agent of the grammatical agreement (controller) [McElree et al., 2003, Lewis and Vasishth, 2005, Van Dyke and Johns, 2012, Wagers et al., 2009, Badecker and Kuminiak, 2007, Dillon et al., 2013, Martin and McElree, 2008].

Importantly, this model is the only model that can explain the phenomenon of grammatical asymmetry [Wagers et al., 2009]. Additionally, unlike the continuous valuation models, this model does not assume separate mechanisms for different linguistic constructions, but rather builds up on an already existing content-addreasable memory mechanism that underlies general purpose working memory [Jonides et al., 2008]. Another advantage of the cue-based model is its ability to account for attraction effects in the case of object-relative clauses, in which the attractor does not appear between the subject-verb dependency (e.g: The boy that the girl likes leaves). Upward percolation cannot account for errors in such sentences

In the account of the cue-based model, the errors stem from a faulty memory access (for example, the head and the attractor noun might be competing for the same memory slot), and the parser temporarily accesses constituents not authorized by the structural operations. That is, attraction effects stem from similarity-based interference at the retrieval stage [Gordon et al.,

2001, Gordon et al., 2002, Tanner et al., 2014]. Furthermore, it has been hypothesized that cue-based retrieval mechanisms might be operating in parallel with predictive mechanisms[Tanner et al., 2014, Tanner and Bulkes, 2015].

Other accounts

Even though the cue-based retrieval model is the dominant model in accounting for attraction phenomena, there exist other frameworks that challenge, or try to augment this account [Jäger et al., 2017].

The cue-retrieval model pinpoints to errors that occur at the retrieval stage of processing, but the encoding process might be an alternative account [Vasishth et al., 2017]. The encoding stage is not taken into account by the cue-based retrieval model.

In a recent study, Villata, Tabor and Julie Frank [Villata et al., 2018] tried to disentangle the two stages reporting data from two self-paced reading experiments of object relative clauses (e.g: The boy that the girl likes leaves) in number and gender agreement in Italian and English. Their results point to an interference profile stemming from both encoding and retrieval processes. Villata and colleagues proposed an augmented version of the state-of-the-art cue-based retrieval computational model (ACT-R, [Lewis and Vasishth, 2005]), that takes the encoding stage into account.

Another recent paper by Christopher Hammerly, Adrian Staub and Brian Dillon [Hammerly et al., 2019] challenges the premise of the cue-retrieval, by providing evidence that the grammaticallity asymmetry is an epiphenomenon of response bias, and that when this bias is regressed out (the response bias was modelled as the rate of evidence accumulation in a diffusion process), the asymmetry disappears (it is worth noting that based on this analysis, the asymmetry was significantly diminished but not entirely vanished). The authors interpreted this results as in line with the family of continuous valuation models, and in particular, the marking and morphing model [Eberhard et al., 2005].

In another recent work, Bojana Ristic, Simona Mancini, Nicola Molinaro & Adrian Staub [Ristic et al., 2021], reported results from two eye-tracking experiments in English and Spanish, that show similarity-based (encoding) interference over the interpolated elements (the elements that carried grammatical number). The authors interpreted these results as evidence of an active maintenance of the number feature of the subject, and therefore as indication in favor of the continuous valuation family models. Notably, the task did not include any comprehension task, and therefore, the grammatical asymmetry cannot be assessed.

Neural correlates of attraction phenomena.

Studies on the neural correlates of interference/attraction phenomena are scarce in general. Additionally, when it comes to classical ERP analyses, Tanner and colleagues have reported results that showcase the extreme sensitivity of language ERP analysis to technical configurations, such as high-pass filtering and averaging.

In particular, Tanner et al. [Tanner et al., 2015] demonstrated that high filtering values usually applied in language ERP studies (e.g: $0.3Hz$ and above) introduced an N400 effect in syntactic violations that was not evident in the unfiltered data. In follow-up works, Tanner and colleagues demonstrated that classical ERP analyses fail to capture individual subject contributions [Tanner et al., 2018]. Using a large cohort of English monolingual participants (N=114), Darren Tanner demonstrated that the individual subject responses in a subject-verb agreement varied in a continuum between the N400 and the P600 component. Importantly, Tanner did not report the LAN or ELAN component, neither at the individual nor at the group level.

Overall, the few ERP studies available point to a response profile in which the main components (N400/P600) are reduced in amplitude in the presence of an attractor [Chen et al., 2007, Kaan, 2002, Severens et al., 2008, Shen et al., 2013, Tanner et al., 2017, Santesteban et al., 2017] and an overall saliency of the attraction effects. That is, the singular head—plural attractor configuration was harder to detect across studies.

Section Summary

Attraction phenomena are bounded by two well showcased asymmetries. The *mismatch* asymmetry that illustrates that attraction effects arise only when the attractor is plural, and the *grammaticallity asymmetry* that depicts that attraction takes place mostly in the agrammatical sentences. Notably, there is a scarcity of studies investigating the neural correlates of attraction.

1.4 Predictive coding as a mechanism of language processing

Two recent accounts hypothesized that the dominant model of attraction (cue-based model) might be operating in parallel with predictive mechanisms [Tanner et al., 2014, Tanner and Bulkes, 2015]. In this section, we review evidence for the role of predictive mechanisms in language processing.

Cognitive systems, whether biological or artificial, are considered fundamentally "prediction engines" [Nave et al., 2020, Clark, 2013], with a growing body of research in cognitive neuroscience claiming that prediction is a "canonical computation" of cognition [Keller and Mrsic-Flogel, 2018].

This is encapsulated in the idea of predictive coding [Friston, 2005] formulated by Karl Friston. Predictive coding assumes that the brain constantly generates and evaluates a mental model of the word. According to this theory, the notion of prediction (expressed via the principle of free energy minimization) is a fundamental property of the human brain[Friston, 2010].)

There is evidence for an active role of prediction in language processing. All the way back to 1984 and the seminal work of Marta Kutas and Steven Hillyard[Kutas and Hillyard, 1984], who showed a modulation of the N400[3] event-related-component (ERP) with the subject's expectancy for the terminal word of the sentence. The authors measured this expectancy via means of "cloze probability" [4]. In a follow-up paper by Delong, Urbach and Kutas [DeLong et al., 2005] participants were asked to provide the optimal continuation for sentences truncated either before the article or the noun. The authors took advantage of the morphological difference of the English indefinite articles ("a" vs "an") and measured scalp potentials of participants during the task. The authors found a strong correlation between the amplitude of the N400 and the cloze probabilities of the nouns and the articles, attributing their results to an anticipatory behavior that the human brain expresses, given that the article 'an' narrows down the number of expected upcoming words. The above-mentioned experiments were based on scalp-potentials, but evidence for the role of prediction in language comes from eye-tracking paradigms as well. Kamide, Altmann and Haywood from the University of York performed a set of three eye-tracking experiments (in English and Japanese) using the 'visual-word' paradigm[5] Their results demonstrated anticipatory eye-movements towards the matching object compared to the unrelated one, providing another piece of evidence for active anticipatory behavior in humans.

[3] Negative scalp polarity that peaks at around 400ms after the onset of the target word

[4] The probability of the target word completing a particular sentence

[5] A setup in which the participants hear utterances while looking at objects which may or may not be associated with the sound. This is a well-studied tool in language research. For a review, see [Huettig et al., 2011].

Evidence of prediction can be traced down to toddlers of two years of age. In a set of picture-priming task experiments performed by Mani, Durrant and Floccia [Mani et al., 2012], the authors provided evidence that word-recognition utilizes a streamlined pipeline of phono-semantically related words, thus pinpointing to an active role of word-prediction even at a very small age. Interestingly, one of the main authors of this paper, co-authored a follow-up review paper entitled *"Is prediction necessary to understand language? Probably not"* [Huettig and Mani, 2016]. The main claim of the authors is that despite the popularity of the predictive coding framework, there is very sparse experimental evidence that the human brain is engaged in predictive coding during language processing, and that may be due to the fact that current (2016) neuroscientific methods are ill-suited to address this question. Additionally, the authors provide a set of five arguments against the necessity of prediction in language comprehension. Amongst them, the fact that prediction is strongly context-dependent and that many experimental set-ups and designs encourage prediction processing. With respect to the argument of context dependency, the authors refer to a previous work of Falk Huettig [Huettig and Guerra, 2015] where the participants were listening to sentences, while presented with pictures of target words. Of these words, only the target was properly gender-marked with the definite determiner of the sentence. The authors offered a preview of the images (the target and the unrelated elements) and varied the speed of this preview, constructing a slow and a fast experimental condition. Huettig and Guerra observed that the prediction effects disappear in the slow condition. These results were interpreted as evidence that prediction mechanisms are not robust to the variability of presentation speed, and thus prediction depends on the context in which a language user is placed. The second argument refers to the fact that many of the typical linguistic and neurolinguistic experiments are prediction-engaging, thus, even thought prediction results are reported, the importance of prediction in language cannot be robustly inferred. Overall, Huettig and Mani [Huettig and Mani, 2016] do not claim that there is no predictive processing in language comprehension and production, but rather question the necessity of prediction in natural language processing.

In a very recent paper that tackled the issues raised by [Huettig and Mani, 2016], Cory Shain et al. [Shain et al., 2020] presented fMRI evidence that addressed the following two questions:

- Does language-specific predictive coding take place in the Language Network (LANG: [Fedorenko et al., 2011, Fedorenko et al., 2012]) or the domain-general, fronto-parietal multiple demand network (MD: [Duncan, 2010]).

- Is language-specific predictive coding sensitive to both structure and surface operations (Hierarchical Syntax Vs n-grams).

This study is unique in two-ways:

- The authors used naturalistic rather than controlled stimuli (the 78 participants of the study listened to stories while scanned). This approach addresses the criticism of using laboratory-controlled settings that might lead to "enhanced" prediction processes [Huettig and Mani, 2016].

- The authors also used a participant-specific localized to identify the distinct networks of interest and a unique, recently developed regression technique which is suited for analyzing BOLD responses from naturalistic stimuli (Continuous Time Deconvolutional Regression-CDR, [Shain and Schuler, 2021], thus partially addressing the issue of the ill-suited neuroscience methods raised by [Huettig and Mani, 2016].

Importantly, the authors contrasted the fit of two models of surprisal[6]. A surface-based, 5-gram model and a hierarchical probabilistic context-free grammar (PCFG), essentially contrasting structural VS n-gram operations.

Crucially, [Shain and Schuler, 2021] found *significant* and *independent* results of both models in the LANG, but not the MD.

The authors interpreted these results as evidence of direct involvement of predictive coding in language processing and as a mechanism operated by a language-specific network. Additionally, language specific predictive-coding appears as a mechanism that operates both at the surface (n-grams) and syntax level, and that different regions of the language network show sensitivity to these distinct mechanisms (temporal and inferior-frontal regions). These results are in contrast to the work mainly driven by Robert Frank, who claims insensibility of the neural system to such effects [Frank and Bod, 2011, Frank et al., 2012, Frank and Christiansen, 2018b]. All in all, the results of [Shain et al., 2020] point to a predictive coding mechanism specialized for language, and sensitive to both surface effects and hierarchical structure operations. The prediction of syntactic structure (as opposed to mere prediction of word-order or semantic associations) is addressed in a very recent review by Fernanda Ferreira and Zhuang Qiu from the University of California, [Ferreira and Qiu, 2021]. The authors claim that syntactic, precedes that of semantic prediction. As a minimal demonstration example, the authors refer to a hypothetical case-study in which only the word "Those" is presented to a subject. Under this scenario, there can be no semantic prediction whatsoever (given the non-existent contextual frame) Nevertheless, the subject can expect that, most probably, a noun in plural will follow immediately (e.g: Those workers). As evidence for prediction of syntactic forms, the authors refer to the seminal paper by Lau et al. [Lau et al., 2006]. In this paper, the authors examined ERPs and in particular the LAN[7] component on the critical word and compared a minimal-pair of the two conditions shown below:

- Although Erica kissed Mary's mother, she did not kiss Dana's *of* the bride.
- Although the bridesmaid, kissed Mary, she did not kiss Dana's *of* the bride.

Both sentences contain a violation and are thus agrammatical. The authors compared the LAN component in the two sentences and found that in the first sentence, the amplitude of the violation-ERP was smaller. The difference between the two sentences is that the first one could have had closure at the last noun (Dana). This fact allows for a so-called elliptical reading (see section **??**). This result is interpreted as a null-form prediction by the syntactic parser, and thus provides evidence of syntactic structure prediction.

In another very recent study, researchers from the Donders Institute in Nijmegen sought to investigate the granularity of linguistic prediction [Heilbron et al., 2021] and addressed the following two questions:

- Is linguistic-prediction ubiquitous?
- What is the level of linguistic prediction?

Similar to the work of [Shain et al., 2020], the authors used naturalistic stimuli. [8] In addition, they utilized a state-of-the-art transformer-based neural network (GPT-2) to quantify

[6] Surprisal is an information-theoretic measure quantifying how unexpected the current word is, given the words that precede it. [Armeni et al., 2017]

[7] Left Anterior Negativity

[8] Two datasets of participants listening to audiobooks. The first (N=19) is a publicly available EEG dataset, whereas the second consisted of 3 participants with custom-made MEG head casts

linguistic predictions in a granular way. The lexical predictions of the GPT-2 were partitioned into distinct linguistic dimensions, which allowed the authors to investigate systematically the level of linguistic prediction (syntactic, semantic & phonemic). Specifically, the syntactic prediction was defined as the conditional probability distribution over parts-of-speech, the semantic prediction as the predicted semantic embedding and the phonemic prediction as the probability of the following phoneme. Using this fine-grained prediction schemas, the authors analyzed the neural responses of the participants using a regression-based deconvolution approach, to study prediction error signals within the continuous recordings. [Heilbron et al., 2021] used three regression models of increasing complexity, and found that the model that included probabilistic information was the best predictor of the neural activity. This result was interpreted as evidence that the brain operates predictive processing on a constant basis. Additionally, the authors reported dissociable patterns of explained variance per participant, and for each prediction level. This was construed as evidence towards feature-specific linguistic prediction, but also, prediction that occurs at both low and high levels of linguistic processing.

Lastly, [Heilbron et al., 2021] decomposed the spatiotemporal dynamics of the syntactic, semantic and phonemic prediction errors, and observed dissociable signatures for each of these levels. Specifically, the syntactic surprise lead to a significant frontal positivity between 200 and 500ms after the onset of the critical word, the semantic effect leaf to a much later positivity, significant between 600 and 1100ms. Finally, the phonemic effect was characterized by a negative, early component, significant between 100 and 500ms. These results were seen as evidence that linguistic predictive coding occurs at multiple networks and at different levels that form a hierarchy of linguistic predictions.

Section Summary

The literature presented in this section draws a clear picture of linguistic predictive coding as an active mechanism of linguistic processing. This mechanism operates at various linguistic levels, from low-level transition probabilities to high-level syntactic operations.

1.5 Non-linguistic sequence processing

1.5.1 A taxonomy sequence processing.

As previously mentioned, language can be decomposed in a sequence of elements, therefore to study language, one should take into account evidence from sequence processing. In a seminal work, Deahene and colleagues[Dehaene et al., 2015] reviewed evidence from sequence processing in humans and other primates, and proposed a taxonomy of five levels of representations of sequences. Of those five levels, four are believed to be shared with other primates. Many other primates represent sequences by their transition and timing (what comes next and at what time: *level*1), they can chunk sequences into groups of objects, discovering that there are so-called words inside the sequence (*level*2), they can decide what comes first, second or third and have number knowledge (*level*3: number sense in other primates), and they can even, in some experiments, understand algebraic patterns.

According to this review, the uniqueness of the human brain can be traced to its ability to represent sequences in the higher level of this codification, that is, representing nested symbolic structures. Notably, W. Tecumseh Fitch, claims that this human tendency to encompass nested symbolic representations expands to high cognitive functions other than language, such as music or mathematics. Fitch terms this ability as "dendrophilia"[9] [Fitch, 2014].

[9] From Greek: δέντρο (tree) + φίλος (friend). Dendrophilia is the love of trees.

1.5.2 The "local-global" paradigm.

The "local-global paradigm" [Bekinschtein et al., 2009a], which is a variant of the auditory oddball paradigm. In the local-global paradigm, participants are presented with sounds in short sequences, and not continuously, 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. 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 [Dehaene et al., 2015].

The neural signatures of these local and global effects differ in several ways. First, the local effect is early (100–200ms) and transient, whereas the global effect requires an additional 100–200ms to rise, and it remains stable [King and Dehaene, 2014]. Second, the local effect is automatic (does not require attention) and unconscious, whereas the global one disappears when participants are not attending or unconscious [Bekinschtein et al., 2009b, Strauss et al., 2015]. Third, while the local effect was traced back to auditory cortices [Pegado et al., 2010], the global one is distributed across the superior temporal sulcus; inferior frontal gyrus, dorsolateral prefrontal, intraparietal, anterior, and posterior cingulate cortices [Uhrig et al., 2014].

1.6 On the disentanglement of structural and linear operations in language comprehension

1.6.1 What this thesis seeks to answer

This thesis seeks to answer an open question in the field of language processing and comprehension: *Is language a phenomenon based on statistical regularities, such as transition probabilities, or is it a phenomenon deeply rooted in a uniquely human ability to represent nested, symbolic structures?* This question is of paramount importance as it addresses the core mechanism of human singularity.

This dichotomy essentially contrasts two distinct mechanisms of language processing. The first mechanism is attributed to probabilistic modeling at the sequence level, and presupposes no structural bias, whereas the second is sensitive to the syntactic structure of the sentence.

The studies mentioned and reviewed above pave the way for this endeavor. We revisit a phenomenon studied since 1924 [Sonnenschein, 1925], armed with predictions from Deep Learning studies and non-linguistic sequence processing. Indeed, evidence from previous studies on human, non-linguistic sequence processing, and artificial neural language models, suggests that these two mechanisms might co-exist, however, for language processing, their isolation remains a challenge.

To disentangle the two mechanisms, we created a new design that utilizes the classical psycholinguistic phenomenon of subject-verb agreement, and in particular, the modulation of this agreement in the presence of a noun that does not intervene structurally with the agreement configuration (hereafter attractor).

To identify neural correlates of these mechanisms, we analyzed data collected from two neuroimaging, and one online-behavioral experiment. Additionally, we employed comparisons between human subjects and deep neural networks, to draw a comparative picture between two systems that are assumed to have a different underlying language apparatus.

Additionally, to address the demonstrated sensitivity of the classical ERP approaches to experimental configurations [Tanner et al., 2015, Tanner et al., 2018], we utilize a multivariate, machine learning analysis approach that provides bigger sensitivity to individual subject contribution, as well as statistical robustness to analyze our data [King and Dehaene, 2014, Dehaene and King, 2016].

Chapter 2

Disentangling Structural and Transition-based Computations during Sentence Processing.

Abstract

Sequence processing in the primate brain is known to rely on multiple mechanisms, including a local mechanism based on transition probabilities, and a more global mechanism based on working memory for long-distance structural regularities. It is debated whether sentence processing also reflects this duality, with one mechanism computing transition probabilities of adjacent elements, and the second computing syntactic tree structures. To disentangle those mechanisms, we examined the brain response to sentence violations in a factorial subject-verb agreement design analogous to the non-linguistic local-global design. In each sentence, one feature of the verb, either animacy or grammatical number, agreed or not with the preceding noun or with a more distant noun (e.g., "The boy near the girls jumps"). We collected electro- and magneto-encephalography signals from 22 human participants, and compared them with the responses of artificial neural networks presented with the same set of stimuli. In the models, transition and structure effects coexisted. However, in humans, we only found evidence for the structural effect. The intervention of an attractor noun with incongruent features did create behavioral interference, but this was due to noun-noun interference in working memory rather than to noun-verb transition probability. Our results point to a major difference between language processing in humans and neural models, as well as between the processing of non-linguistic and linguistic sequences in humans.

2.1 Introduction

Sequence processing was conjectured to take place across multiple levels in the brain, from low-level, transition-based processing, to high-level processing of tree-like representations [Dehaene et al., 2015]. According to this view, incoming sequences are internally represented across a hierarchy of five distinct levels, each with corresponding cerebral mechanisms and specific properties: (1) transitions and timing knowledge, (2) chunking, (3) ordinal knowledge, (4) algebraic

patterns, and (5) nested tree structures. While the lowest level in this hierarchy is assumed to pertain to various types of sequence processing, and to be found in animals, the highest level, that of tree-structure representations, is considered to be specific to human language. Following this multi-representational view, here, we study whether sentences in language are processed at multiple levels, involving both structural and transition-based computations, and whether distinct neural mechanisms underlie different levels of processing.

The multi-representational view to sequence processing was based on evidence coming from the processing of non-linguistic sequences. In several studies, it was shown that separate brain regions are involved in the processing of different levels of encoding of incoming sequences. Specifically, using the Local-Global Paradigm, a variant of the oddball auditory paradigm, [Bekinschtein et al., 2009b] showed that transition-based (local) processing can be distinguished from that of chunking (global) and from possibly higher levels, eliciting signals in disjoint brain regions [El Karoui et al., 2015, King and Dehaene, 2014, Strauss et al., 2015]. Extrapolating these findings to language, we ask whether during sentence processing, distinct neural mechanisms might underlie transition-based (local) and structure-based (global) processing of word sequences in the human brain.

Recent studies on artificial Neural Language Models (NLMs) provide further support to the multi-representational view. In these studies, it was shown that distinct neural mechanisms, which spontaneously emerge in the models during training, underlie two types of computations during language processing [Lakretz et al., 2019a, Lakretz et al., 2021c]. One type of mechanism was shown to be sensitive to the latent structure of the sentence, and to be carried by 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 different set of units, termed 'Short-range' units. While the short-range units carry predictions about upcoming words based on 'low-level' properties of the sentence, the long-range units generate predictions based on the hierarchical structure of the sentence. Predictions in the models about upcoming words were therefore shown to be composed of processes occurring at distinct levels and to be either structure sensitive or structure agnostic.

Suppose that the human brain also contains such two distinct types of neurons, in low- and high-level regions, which are either structure-sensitive or structure agnostic. Then, during sentence processing, we might observe neural activity arising from such distinct neural mechanisms, similarly to what has been observed for sequence processing of non-linguistic stimuli. To test this hypothesis, we introduce a new design, akin to the local-global paradigm, which directly contrasts structure-agnostic (local) and structure-sensitive (global) computations during sentence processing. We recorded neural activity from both human participants (n=22, magnetoencephalography; MEG) and neural language models (n=20) and studied whether the neural signatures of the two levels can be disentangled. We used temporally resolved decoding techniques [King and Dehaene, 2014], which we applied to neural data from both humans and neural language models. In the models, we found evidence for two distinct neural effects, which correspond to the two types of processing, corroborating previous results in NLMs. With humans, however, only a structural main effect was found, whereas decoding of the transition effect remained at chance level. Yet, the structural effect was modulated by the transition one, as revealed by an interaction analysis, providing indirect evidence for the existence of also word-transition computations during language processing. However, this discrepancy in effect size suggests that unlike low-level processing of auditory stimuli, in sentence processing, once word sequences enter the language system they are dominated by structure-based processing and are largely robust to transition effects.

Finally, the global effect was further modulated by grammatical number, which provides the evidence for neural correlates of the markedness effect [Bock and Miller, 1991], so far mainly

described in behavioral data.

2.1.1 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 [Bekinschtein et al., 2009b], 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 [Bekinschtein et al., 2009b]. 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 [Dehaene et al., 2015]. The neural signatures of the local and global effects further differ in several ways. First, the local effect is early (100–200 ms) and transient, whereas the global effect requires an additional 100–200 ms to rise, and it remains stable [King and Dehaene, 2014]. Second, the local effect is automatic (does not require attention) and unconscious, whereas the global one disappears when participants are not attending or unconscious [Bekinschtein et al., 2009b, Strauss et al., 2015]. Third, while the local effect was traced back to auditory cortices [Pegado et al., 2010], the global one is distributed across the superior temporal sulcus; inferior frontal gyrus, dorsolateral prefrontal, intraparietal, anterior, and posterior cingulate cortices [Uhrig et al., 2014]. 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 hypothesized 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 [Franck et al., 2007]. 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 nested prepositional phrase (Nested-PP; Figure 2.1A):

(1) **Nested-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).

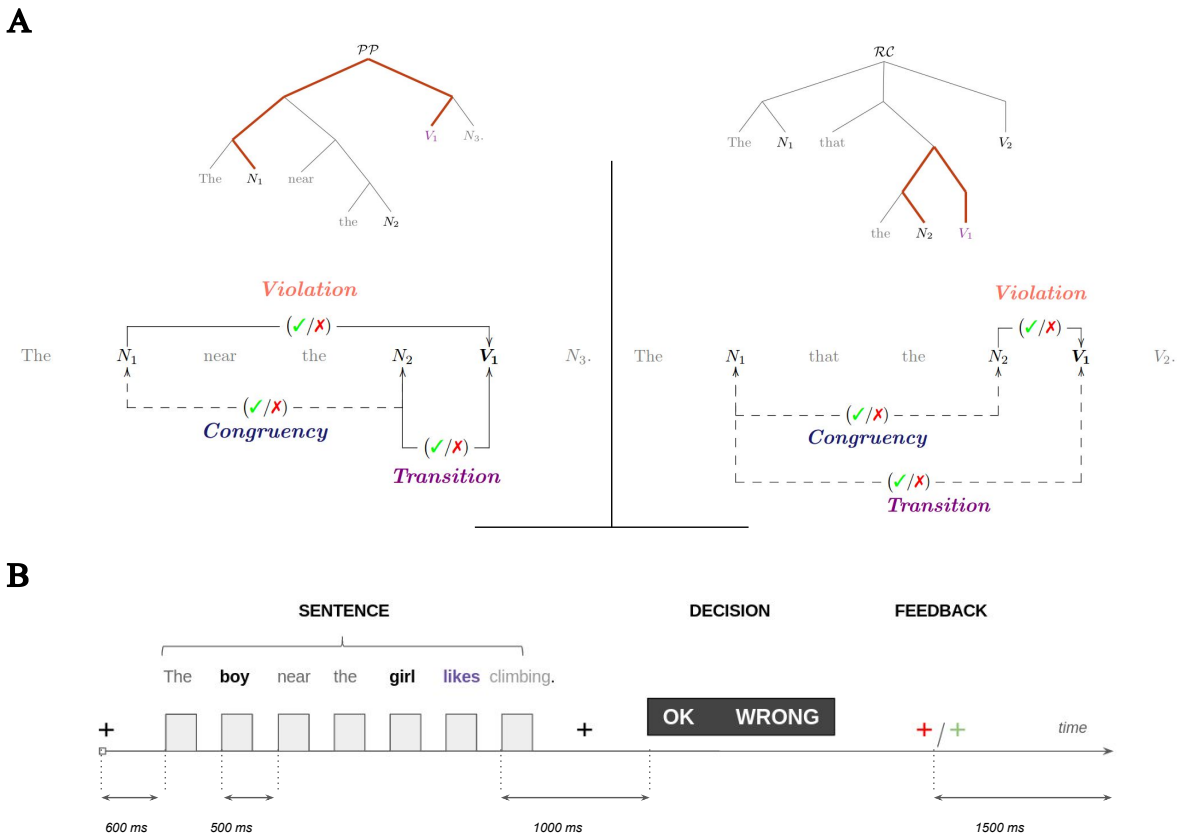


Figure 2.1: **Structural vs. linear intervention in sentence processing—experimental design and paradigm.** To disentangle two possible types of processing during sentence comprehension, the experimental design contrast: (i) a structural dependency between a target verb and a noun, which either holds 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: Violation effect (orange), which depends on the structural 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, with respect to either grammatical number or animacy. Congruency effect, which refers to the (mis)match between the two nouns; In the left construction, the violation (structural) effect is long-range and the transition (linear) one is short-range. On the right construction, it's 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.

Construction	Feature	Template	Example	Violation	Congruency
Nested-PP	Number	*Det* N_1 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
Nested-ObjRC	Number	*Det* N_1 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 leave.	Yes	Yes
Nested-PP	Animacy	*Det* N_1 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

Figure 2.2: **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). Based on the main effects of Violation, Congruency, and Transition, the design can be interpreted as a 3×3 factorial design.

These two dimensions span the two-by-two design of the paradigm, and its corresponding two main effects are defined as follows (Figure 2.1): 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 [Osterhout and Holcomb, 1992, Osterhout and Mobley, 1995]. 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 example, 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 [Friston, 2005, Friston et al., 2021], 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 (see also [Kuperberg and Jaeger, 2016], for a multi-representational hierarchical generative approach to language comprehension).

2.1.2 Decoupling syntactic dependency and linear proximity

Notice that in the construction with a nested 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 nested PP by a nested object-relative clause (ObjRC; Figure 2.1, Table 2.2):

(2) **Nested-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 nested-PP but not in the nested-ObjRC case.

2.1.3 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 [Friederici et al., 1999], semantic violations were found to elicit an earlier negative response N400 [Hagoort, 2003]. 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) **Nested-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 2.2 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 a nested object-relative clause (3) and semantic violations.

2.2 Methods

Participants 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 for their participation. Prior to the participation, the subjects had to perform an online sentence reading task (Dialang reading task). 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 violation to the grammaticality of the task 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 2.1B). 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 multiplications of the video projector refresh rate ($60Hz$).

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

For each of the three constructions (Table 2.2), 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($N = 15$), ICM ($N = 7$). 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, EEGrecording 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 [Gramfort et al., 2013, Jas et al., 2018]. To compensate for head movement and reduce non-biological noise, the MEG data were Maxwell-filtered [Taulu et al., 2004] using the implementation of Maxwell filtering in MNE Python. The bad EEG sensors were interpolated using the spherical spline method [Perrin et al., 1989] 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 [Gramfort et al., 2013]. Finally, the 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.

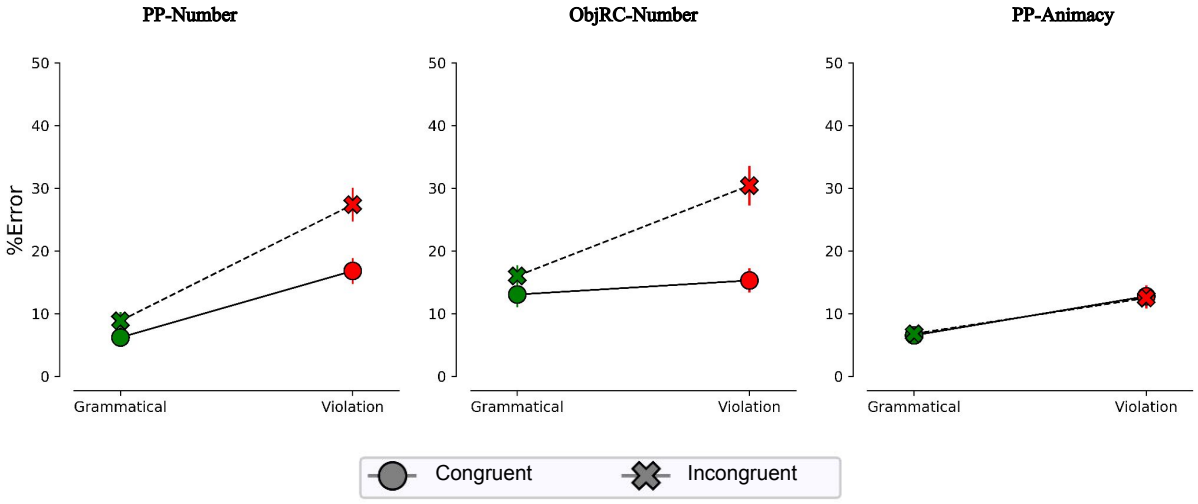


Figure 2.3: Behavioral results. Interaction plots for the Violation and Congruency effects (N=22). The main effects for Violation and Congruency as well as their interaction, are significant in the number constructions ($p < 0.05$). In the animacy condition, only the main effect of Violation is significant. The error bars indicate the standard error of the mean (SEM) calculated across participants.

Decoding Analyses We used a temporally defined decoding approach to classify neural activity from two conditions at the trial level [King and Dehaene, 2014, Dehaene and King, 2016]. These analyses were implemented in MNE-python version 0.16 [Gramfort et al., 2013]. Prior to model fitting, the data were standardized using the Scikit-Learn package [Pedregosa et al., 2011]. 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 [Gramfort et al., 2013]. The statistical significance of the decoding performance over time was evaluated and corrected for multiple comparisons using a cluster-based permutation approach [Maris and Oostenveld, 2007], using a total of 1000 permutations. The significance threshold (alpha level) for all analyses was set to 0.05.

2.3 Results

In this project, we sought to disentangle structural from transition (surface statistics) effects. To analyze the data, we parsed the factorial design based on two main factors and their corresponding interaction. We first introduce behavioral results from the experiment, based on the performance of the subjects in a forced-choice, violation-detection task (Figure 2.1C). 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.

2.3.1 Behavioral Results

Figure 2.3 shows the mean error rates across all participants for the three main constructions. We present the error-rates with respect to the two main factors of the design: violation and congruency.

We first tested whether the structural effect (violation – Figure 2.1C), is observed at the behavioral level. Indeed, this effect was significant across all constructions. This effect indicates that participants made more errors in detecting a violation, compared to an affirming that a sentence was grammatical, and coincides with the well reported phenomenon of grammatical illusions [Wagers et al., 2009]. The effect was stronger for PP-Number (Violation: 22.11.78%; Grammatical:7.530.92%; $F(21) = 56.94$, $p < 1e-10$), followed by PP-Animacy (Violation: 12.631.2%; Grammatical: 6.660.81%; $F(21) = 16.85$, $p < 1e-04$) and finally ObjRC-Number (Violation: 22.862.01%; Grammatical:14.521.32%; $F(21) = 13.25$, $p < 1e-03$).

We then examined the modulation of the error-rate by the congruency effect. Notably, this effect was significant only for the number feature. In other words, 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 in animacy. Importantly, the effect was weaker compared to that of violation for both constructions (PP & ObjRC). We also observed within-construction differences. In particular, the effect was weaker in the PP-Number construction, (Congruent: 11.51.31%; Incongruent: 18.121.81%, $F(21) = 11.77$, $p < 1e-03$) compared to the ObjRC (Congruent:14.161.4%; Incongruent: $23.22, 1.4\%$; $F(21) = 15.67$, $p < 1e-03$)

The congruency effect stems from a factorial manipulation of features across the two nouns of the sentence. We then sought to investigate the effect of a mismatch between the non-structurally intervening noun, and the target verb. This effect can be realized as the interaction of the main factors of congruency and violation, and in the case of the PP construction, it corresponds to a bigram transition effect between the non-structurally intervening attractor and the verb. Notably, as with the congruency effect, this effect was significant only in the case of the number feature.

In both constructions, the interaction was weaker compared to the violation and congruency factors and for the long-range dependency (Nested-PP-number), it was marginally significant ($F(21) = 3.99, p = 0.04$). It is important to note, that in the ObjRC construction, this interaction reflects the interference produced by a distant element (N1) that disagrees in grammatical number with the target verb. Simply put, this effect does not coincide with a transition phenomenon, as it does for the PP-Number construction. It is worth noting, then, that for this construction, we observed a stronger interaction effect compared to the Nested-PP condition, nevertheless, still a weaker effect compared to the effects of congruency and violation ($F(21) = 6.2, p = 0.01$).

In summary, at the behavioral level we observed a clear modulation of the error rate by the structural effect of violation, for both constructions and features. Additionally, we reported a modulation of the error-rate by effects stemming from the influence of a non-structurally intervening noun. Notably, we reported interference effects in the case of a non-embedded attractor (ObjRC construction), in agreement with previous studies [Wagers et al., 2009]. Importantly, the latter effects were weaker compared to the structural effect of violation.

Finally, we detected a significant construction effect. Overall, participants made more errors on sentences with a nested object relative clause compared to a nested prepositional phrase ($PP: 3.052.9\%$; $ObjRC = 4.643.31\%$, $F(21) = 15.82$, $p < 1e-03$).

Construction	Feature	Violation		Congruency		Violation:Congruency (Interaction)	
		F	p-value	F	p-value	F	p-value
Nested-PP	Number	56.94	<e-11	11.77	<e-03	3.99	<e-01
Nested-ObjRC	Number	13.25	<e-04	15.67	<e-03	6.2	<e-01
Nested-PP	Animacy	16.85	<e-04	n.s	n.s	n.s	n.s

Figure 2.4: Three one-way between subjects ANOVAs were conducted to compare the effects of congruency, violation, and their interaction on the Error Rate

2.3.2 Structural but not Transition Effects are Decodable in Human Data.

Subsequently, we tested whether these effects are traceable in the neural data. We tested this in both neural-network language models and humans. We presented the same stimuli to human participants and to the models, and recorded network activity after the presentation of each word. For humans, neural activity was recorded with a magnetoencephalography (MEG) machine. For models, 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 2.5 shows the decodability of the main effects for both the artificial (panel A) and human data (panel B).

For the models (Figure 2.5A), for all three constructions, all effects were decodable with high performance, measured in terms of the Area Under of Curve (AUC). The violation 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 congruency 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 models, only the structural effect of violation was decodable. The onset and the peak decodability of the violation effect varied across constructions. The effect becomes significant first, for the Nested-ObjRC-number ($t = 300ms$), then for the Nested-PP-Number ($530ms$), and lastly for the Nested-PP-animacy ($760ms$). The significance of the decodability was calculated based on cluster-based permutation testing [Maris and Oostenveld, 2007]; Methods). The decodability of the violation effect reached its highest value for the Nested-ObjRC-number construction ($AUC : 0.61$), followed by the Nested-PP-Number ($AUC : 0.58$) and finally by Nested-PP-animacy ($AUC : 0.55$). We observed a discrepancy between the behavioral results, the decoding from the neural networks, 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 attractor were significant (but weaker) only for the number feature. In the networks, we observed the same sensitivity across all three main factors, namely, clear effects arising from the attractor for both the animacy and the number feature. Lastly, in the human neural data, only the structural effect was decodable, whereas the performance of the attraction effects remained at chance level. It is worth noting, that the simulations we run in the language models might be considered as noiseless "recordings". Therefore, one reason that we failed to detect the attraction effects at the neural level, might

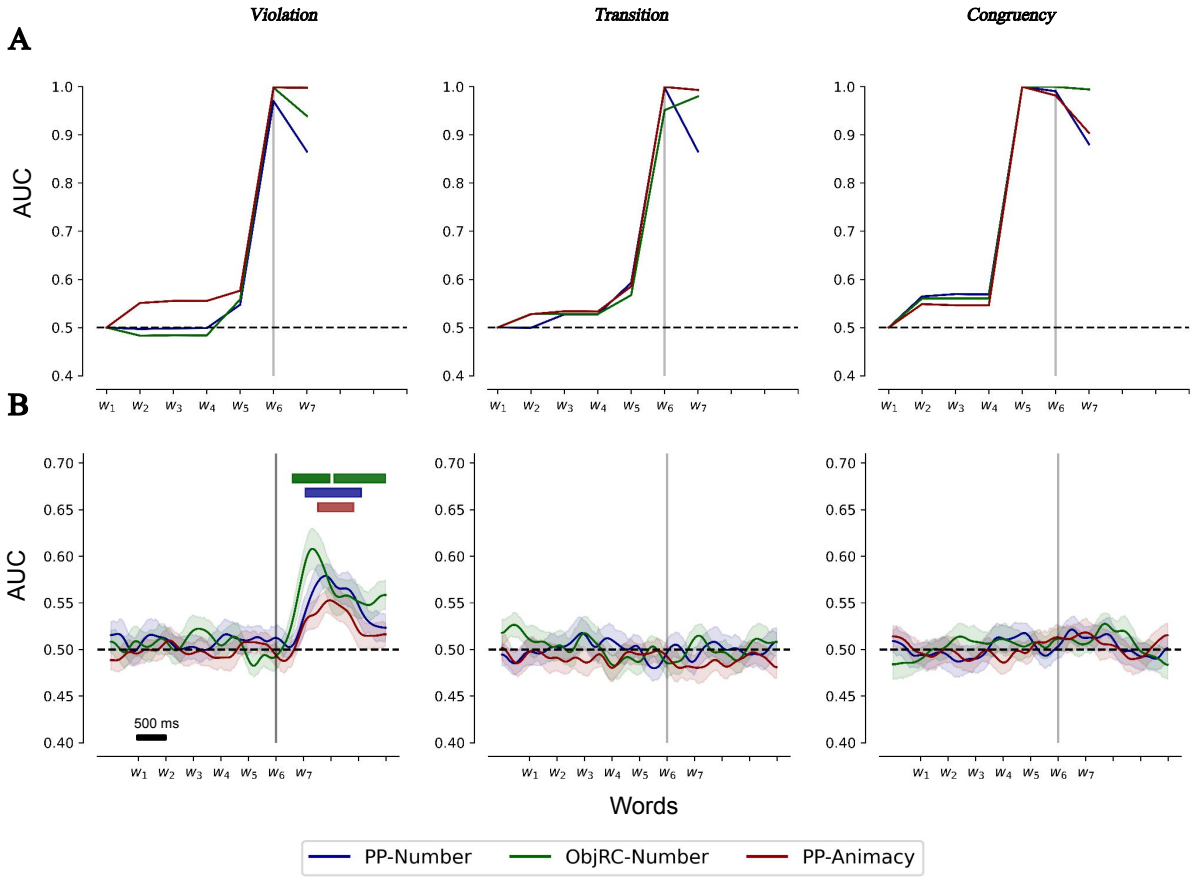


Figure 2.5: **Structural but not linear effects are decodable in human data.** In contrast, all effects are decodable in LSTM activations. (A) Decoding of the main effects originating from the activations of an LSTM architecture. All main effects are decodable. (B) Neural decoding of the three main effects using all sensor types (magnetometers, gradiometers and eeg). A different decoder was evaluated per time-point and modifier. The evaluation metric is the Area Under the Curve (AUC). Only the main effect of Violation (A) is decodable. The dotted lines indicate statistically significant time intervals ($p < 0.05$; corrected—spatio-temporal clustering permutation test). The decoding for the main effects of Transition (B) and congruency (C) remained at chance level until the end of the time of interest. Results shown for correct responses (See S3 for all responses). Data smoothed with a 100ms moving Gaussian kernel for visualization purposes.

be that their size is significantly smaller, and the signal-to-noise ratio (SNR) of the M/EEG does not allow for their detection. Given the evidence for the existence of these effects from the behavioral analysis, we sought to analyze the data employing a second-order analysis.

2.3.3 The influence of the attractor on the structural effect of violation.

We aimed to improve the SNR by increasing the number of samples on which the decoder is trained. We therefore examined the modulation of violation by both the congruency (Figure 2.6A) and transition effects (Figure 2.6B), employing a second order approach. For this, and for each construction, we trained a linear binary classifier on the violation effect, and then, separately, tested it on different conditions in the test data. For example, for modulation of violation by congruency (Figure 2.6A), 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 lines). The same approach was taken to examine the modulation by transition. At test time, the classifier was tested on trials for which the second noun agreed (match) or disagreed (mismatch) in feature with the target word.

Figure 2.6 shows that for Nested-PP-Number, the violation effect is modulated by both congruency and linear interference. The difference between the congruent and incongruent trials became significant at $580ms$ after the onset of the target, and was sustained for $200ms$ (Figure 2.6A). In contrast, the difference emerging from the transition effect became significant much earlier ($300ms$) and was sustained for another $300ms$.

For Nested-objRC-number, violation was only modulated by congruency. This modulation became significant later compared to the one for the Nested-PP-Number construction, starting at $780ms$ and up until $980ms$ after the target onset. It's important to note that the linear interference effect did not modulate violation for this construction. In this case, the effect is not defined as a transition probability between two adjacent elements, but rather as an intervention caused by the first noun of the sentence. Finally, for Nested-PP-animacy, violation is neither modulated by linear interference nor by congruency, in complete agreement with the behavioral results.

2.3.4 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 [Bock and Miller, 1991]. 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: [Bock and Miller, 1991, Eberhard, 1997, Wagers et al., 2009] ; Italian: [**?**] Spanish: [Bock et al., 2012, Lago et al., 2015] French: [Franck et al., 2002] Russian: [Lorimor et al., 2008].

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 congruency effect modulated the effect of violation (Figure 2.6B). To investigate the markedness effect, we took this analysis one step further. A linear binary classifier was trained on the violation effect, and then at testing time, asked to classify

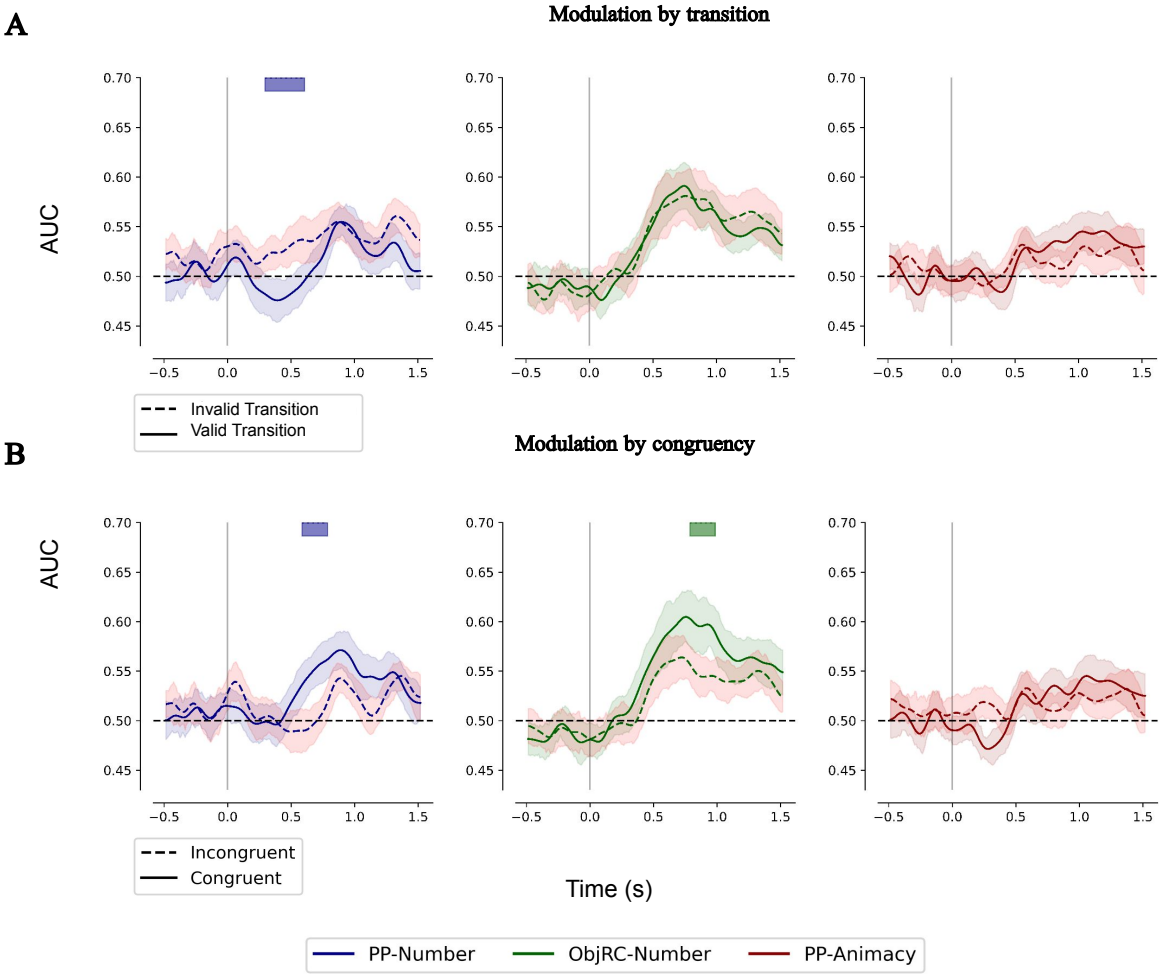


Figure 2.6: **Modulation of the structural effect by transition and congruency.** A classifier was trained on the main effect of violation per modifier, and subsequently tested on the violation effect when (A) contrasting the local standard (continuous line) and local deviant (dashed line) trials, (B) contrasting the congruent (continuous line) and incongruent (dashed line) trials. The performance was evaluated using the AUC metric.

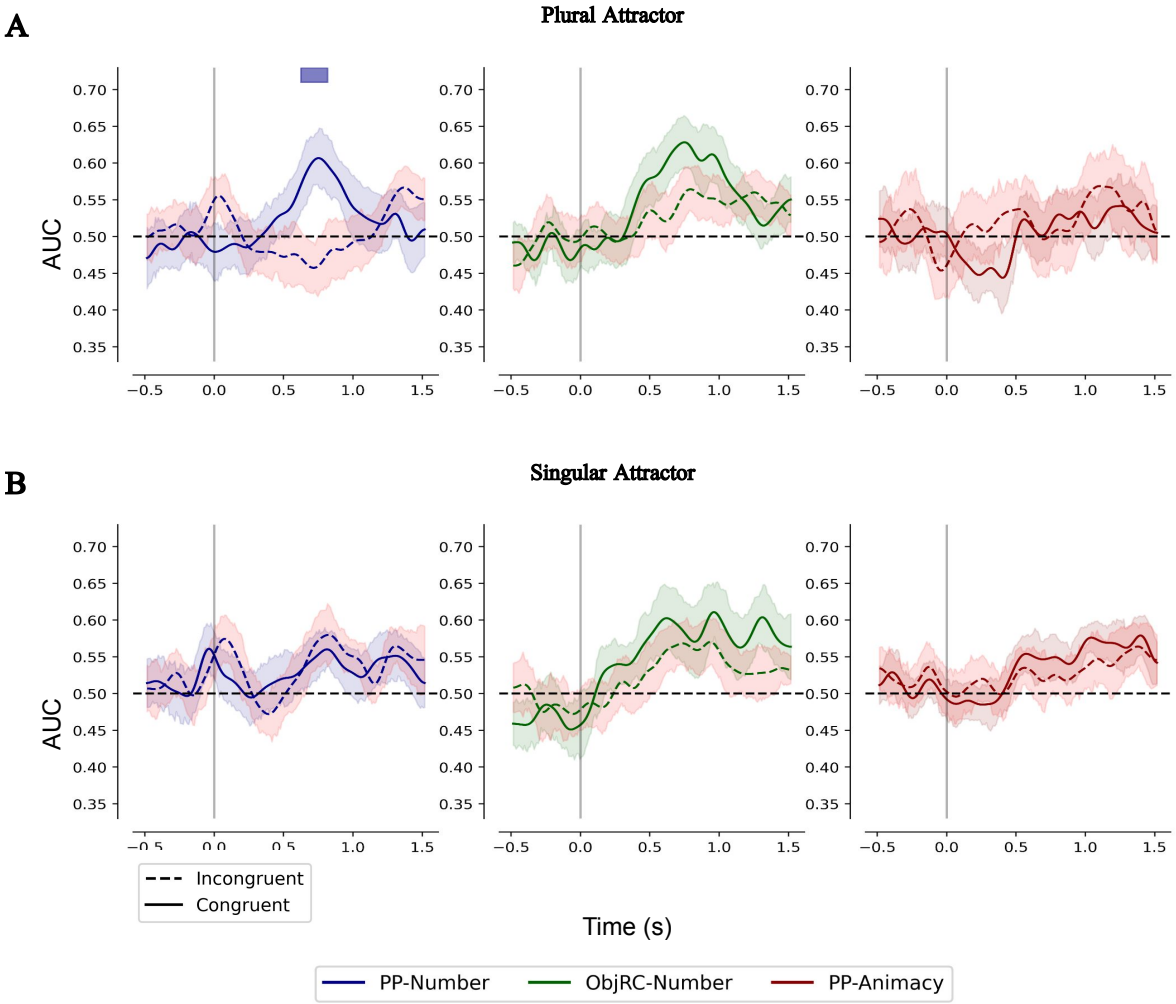


Figure 2.7: **The neural correlate of the markedness effect.** A classifier was trained on the main effect of violation per modifier, and subsequently tested on the violation effect of trials which were split for congruency and attractor number. There is no significance different between the congruent and the incongruent trials for any of the modifiers in the case of the singular attractor (A). In contrast, there is a statistically significant difference($p < 0.05$; corrected—cluster based permutation test) between the two in the case of the plural attractor.

trials based both on their congruency and attractor number (e.g: congruent-singular attractor vs incongruent-singular attractor).

When examining the modulation of violation by congruency only for the singular attractor, we observed no difference between the congruent and incongruent trials, for none of the three constructions. On the contrary, 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 the target word. Figure 2.7 summarizes the neural correlates of the markedness effect. This significance was calculated based on cluster-based permutation testing using a third-order interaction analysis where the three factors of the analysis were the grammatical number, the congruency and the construction. The significance of the decodability for this interaction was calculated as the difference of the differences of the three factors against the chance level. [Maris and Oostenveld, 2007], 2007; Methods).

2.4 Discussion

Evidence from animals, humans, and neural language models suggests that non-linguistic sequence processing takes place across multiple levels, simultaneously, 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 [Bekinschtein et al., 2009b, Dehaene et al., 2015, Dehaene and King, 2016]. Likewise, analyses of artificial neural networks have concluded that both mechanisms contribute to next-word prediction in AI language models [Lakretz et al., 2019a]. In the present study, we tested whether this is also during human language processing: could 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 transition- 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 predictions 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 feature 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 con-

sistent with memory-based models of sentence processing [Lewis and Vasishth, 2005] for which it was suggested that morphosyntactic processing is relatively 'fragile' compared to processing of animacy [Stoops and Christianson, 2017].

Memory-based models of sentence processing [McElree et al., 2003, Lewis and Vasishth, 2005, Van Dyke and Johns, 2012, Wagers et al., 2009, Badecker and Kuminiak, 2007, Dillon et al., 2013, Martin and McElree, 2008] 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 [Stoops and Christianson, 2017], 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 congruency (Figure 2.3, Table 2.4). That is, an incongruency between the main and local noun elicited more errors in ungrammatical compared to grammatical sentences, more 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 [Wagers et al., 2009, Lago et al., 2015]. In these studies, it was shown that agreement attraction facilitated the processing of ungrammatical but not grammatical sentences, in the case of incogruent nouns.

Memory-based models of sentence processing [Lewis and Vasishth, 2005, Wagers et al., 2009] 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 congruency, 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 agrammatical, either because of the correct local noun-to-verb transition ("girls are"), or because the incongruent noun-to-noun relationship (the plural of "girls" would interference 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.

To analyze the neural data, we adopted a multivariate decoding approach, 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 (e.g., [King and Dehaene, 2014, King et al., 2018].

In the models, based on previous findings about two neural mechanisms underlying morphosyntactic processing [Lakretz et al., 2019a], 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 num-

ber 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 effects for both structural, transition and congruency effects (Figure 2.5). For the structural and transition effects, maximal decoding was reached at the verb onset, whereas for the congruency effect it occurred on the preceding noun, since information about congruency is already available at this point.

Crucially, the human data contrasted sharply with the predictions from the models. For all three constructions, no transition-based effect was ever observed, and only the structural effect was significant (Figure 2.5). Furthermore, the structural effect was modulated by noun-noun congruency (Figure2.6A), thus providing an explanation for the findings from the behavioral data.

Taken together, our results suggest a substantial difference between how humans and models process language, and importantly, a difference between how non-linguistic sequences and language are processed in humans. Specifically, in the case of the local-global paradigm with sequences of auditory tones, transition effects are easily detectable, and it is rather the 'global' effect, which is more difficult to identify in the neural data. For language processing, our results point to the complete opposite – while the structural effect is large and easily detectable, the transition effect is barely detectable or non-existent. This shows that 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.

Chapter 3

Neural correlates of subject-verb agreement resolution: A multiple stimulus onset asynchrony study in French.

Abstract

In this study, we investigated an open question in language comprehension, and sought to identify neural correlates of two discrete mechanisms of language processing. The first mechanism (hereafter *structural* is grounded in linguistic theory and pre-assumes a hierarchical encoding of sentences, whereas the second is attributed to probabilistic modeling, and presupposes no structural bias (hereafter *linear*. To tackle this question, we used multivariate analysis to analyze behavioral and neural data (EEG & MEG recordings, $N = 20$) originating from a forced-choice, violation-detection task. In our previous work, we failed to detect direct neural signatures of a purely linear mechanism originating from operations that can be attributed to transition probabilities between non-structurally adjacent words.

We hypothesized that three factors might be responsible. First, the morphological complexity of English might not have been sufficiently complex to provoke such effects, given that the inflectional difference of the singular and the plural tense is mostly based on a single letter. Second, we operated under the hypothesis, that transition-based effects might be too fast of a process and therefore dissolved in our original slow SOA settings. Third, we framed the hypothesis that the subjects employed task-resolution strategies due to the lack of filler trials. To that end, we launched a new M/EEG experiment using French, a language with richer inflectional morphology compared to English, utilizing a carefully selected lexicon. In this study, we expanded our previous design, by introducing a modulation of the stimulus onset asynchrony ($125, 250, 375, 500ms$) to address the working hypothesis that purely-transition effects might not be detectable in slow SOA settings. Additionally, we included trials where violations occurred in different places, compared to those of our canonical conditions

Compatible with our previous results, we managed to decode the effect attributed to the structural mechanism. When filtering for the correctness of the responses, we traced effects stemming from a noun-noun dependency. Importantly, these effects were consistently late, surfacing after the onset of the structural effect.

Notably, we could not decode any effects attributed to transition-probabilities between the attractor and the target verb. Thus, in agreement with our previous work, we could not trace neural correlates of transition probabilities between linearly-but, not structurally-adjacent words. In contrast, the structural effect was significantly detectable. Our results replicate the findings and conclusions of our previous work. Language processing is driven by structure-based computations and is robust to transition probabilities between non-structurally adjacent words.

3.1 Introduction

Language processing appears as a trivial task accomplished within milliseconds by the neural system. Nevertheless, the underlying mechanisms of language comprehension are not entirely unriddled.

Reviewing evidence from sequence processing in humans and other primates, [Dehaene et al., 2015] postulated a taxonomy of sequence representation. According to this codification, the lowest level of sequence-processing pertains to effects of transition probabilities between adjacent elements, a processing that can be attributed purely to Markov chain operations and stems from the predictive coding theory [Friston, 2005, Friston et al., 2021]. In contrast, the highest level of sequence processing is the ability to manipulate nested tree structures. This project seeks to answer an open question in the field of language processing and comprehension, originating from the hierarchy of sequence representations [Dehaene et al., 2015] - *Does language processing engage both simple statistical regularities, such as transition probabilities, and structural operations, deeply rotted in complex tree structures?*

To that end, we focus on the phenomenon of subject-verb number agreement. This grammatical aspect refers to the ability to encode and store the feature of grammatical number information of the subject across many elements of a sentence, until the target (verb) is reached [Molinaro et al., 2011a].

On the one hand, linguistic regularities, such as feature agreement, have been described by linguists under the assumption of structural computations. Symbolic structures, often termed as syntactic trees, can describe the sentences, and the structure of the trees can be used to infer the agreement configuration [Franck et al., 2007, Franck et al., 2006, Franck et al., 2010]. Structural operations are considered a hallmark of language processing, rooted in an innate human ability for recursion [Chomsky, 1957, Rizzi, 2004, Cinque and Rizzi, 2010, Dehaene et al., 2015].

On the other hand, language processing utilizes prediction at multiple levels [Heilbron et al., 2021]. For example, probabilistic processing is summoned to resolve ambiguity when multiple structures are available in a given input [Levy, 2008]. In general, probabilistic processing is considered a mechanism that the parser employs to process linguistic input in the presence of noise [Gibson et al., 2013]. The debate between statistical, word-order processing (hereafter *linear operations* and symbolic, *structural operations* is a topic that often stirs up heated controversy [Frank and Christiansen, 2018a].

In previous work (Chapter 4 we sought to identify distinct, incontrovertible signatures of both types of processing in a combined M/EEG[1], RSVP[2] experiment in English, with an SOA[3] of 500ms. Whereas, we were able to successfully decode neural activity attributed to structural processing, we failed to identify any effect ascribed purely to linear operations. We hypothesized that three factors might be responsible for the absence of word-order effects. First, it might be that phenomena stemming from transition probabilities are extremely fast processes, and therefore our relatively slow SOA failed to capture them. Second, the lack of filler-trials[4] might have resulted in the development of task-resolving strategies (i.e., encoding the number of the main subject and wait until the verb is presented, ignoring all that appears in between, including the attractor.) Third, the morphological complexity of English might not be sufficient to elicit strong word-order effects in a subject-verb agreement M/EEG setting, given that the covariation of inflection morphology in English is mostly attributed to an addition or subtraction of a single letter ("s"). Finally, we conjectured that the Signal-to-Noise-Ratio (SNR) of our non-invasive

[1] Magnetoencephalography & Electroencephalography
[2] Rapid Serial Visual Presentation
[3] Stimulus Onset Asynchrony
[4] Sentences where the violation occurs in positions other than the target verb.

experimental setting might have been low enough, for these effects to remain undetectable. In this project, we address the abovementioned issues in a two-fold manner. Inspired by the work of [Vagharchakian et al., 2012], we utilize a new M/EEG design with a multiple of SOAs in French, where we carefully constructed the lexicon to maximize the inflectional difference of the verb conjugation (see section 3.2. Additionally, to tackle the probable cause of the low SNR, we run the English version of the experiment in patients implanted with intracranial recording sites. In this project, we present evidence from a complete M/EEG cohort and preliminary, single-subject, intracranial results.

3.2 Methods & Materials

Participants

M/EEG

A total of 23 participants were recruited. Two participants were exlcuded due to a technical malfunction, and one due to overall bad performance (chance level). Therefore, a total of 20 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 for their participation. The procedure and the consent were approved by the local ethical committee (Université Paris-Saclay, ref. CER-Paris-Saclay-2019-063). Recordings took place at the NeuroSpin research center, in Paris-Saclay, France. After the end of the experiment, the participants were asked to provide feedback[5] and answer a few questions regarding the experiment.

Intracranial Recordings

A total of 4 participants participated in the intracranial recording experiments after written informed consent was obtained. All experimental procedures were reviewed and approved by the Committee for the Protection of Human Subjects (CPHS) of the University of Texas Health Science Center at Houston. Inclusion criteria for this study were that the participants were English native speakers, left hemisphere dominant for language and did not have significant additional neurological history (for example, previous resections, MR imaging abnormalities such as malformations or hypoplasia). Recordings took place at the Memorial Hermann Hospital, in Houston, Texas, USA, under the supervision of Dr.Nitin Tandon.

Experimental Paradigms

In both experiments, participants undertook a rapid serial visual presentation, forced, binary-choice, violation-detection task. To verify that participants understood the task, prior to recording, they went through a short training phase. A $600ms$ fixation-cross interval preceded the onset of the first word (Figure 4.1C). All sentences had the same length. In both experimental designs, a decision panel with the words "OK" and "WRONG" appeared on the screen $1250ms$ after the onset of the last word.

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

[5] https://tinyurl.com/languageexpfeedback

Construction	Template	Examples	Violation	Congruency
$PP - Number$	Det N_1 P det N_2 V_1 N_3	The doctor near the nurse fears the dog.	No	Yes
		The doctor near the nurses fears the dog.	No	No
		The doctor near the nurses fear the dog.	Yes	No
		The doctor near the nurse fear the dog.	Yes	Yes
$ObjRC - Number$	Det N_1 Adv det N_2 V_1 V_2	The doctor that the nurse fears operates tomorrow.	No	Yes
		The doctor that the nurses fear operates tomorrow.	No	No
		The doctor that the nurses fears operates tomorrow.	Yes	No
		The doctor that the nurse fear operates tomorrow.	Yes	Yes
$PP - Filler$	Det N_1 P det N_2 V_1 N_3	The doctor near the sneezes fears the dog.	Yes	-
$ObjRC - Filler$	Det N_1 Adv det N_2 V_1 V_2	The doctor that the nurse fears farmer tomorrow.	Yes	-

Table 3.1: Prototypical sentences used in the French M/EEG experiment, translated in English. For a set of French sentences, see table 3.3

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 multiplications of the video projector refresh rate (for the M/EEG experiment) and the personal computer for the intracranial experiment. Both refresh rates were 60 Hz.

The only difference between the two designs, is that the length of the English sentences was seven words, whereas the French sentences had a length of eight words. Tables 3.1 & 3.4 provide a set of prototypical sentences used in the two experiments.

Stimuli

In both designs, we used two linguistic structures, a Nested Prepositional Phrase (PP) and a Nested Object Relative Clause (ObjRC). Unlike our previous work (Chapter 4, we only focused on the feature of grammatical number. This configuration led to the conception of two constructions (PP-Number & ObjRC-Number). Additionally, we introduced the inclusion of filler trials. We used both, Part of Speech (PoS), and grammatical number fillers. We aimed at introducing violations in locations other than those occurring during the presentation of the sentences of interest. Therefore, we embedded violations in the interior part of the PP (i.e: $N_1 \rightarrow verb$) and the exterior part of the ObjRC (i.e: $V_2 \rightarrow noun$). Table 3.3[6] summarizes the designs and presents prototypical sentences used in the experiments. In the French, M/EEG experiment, filler trials occupied 11% of each run, whereas in the intracranial experiment 10%.

French Stimuli

Each construction was equally presented across the four SOAs for a total of eight runs. For each of the two constructions, we generated 16 stimuli per run, leading to a total of 128 trials per SOA, for each subject. 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.

A custom-made web-scraper was used to download the whole list of second and third group French verbs from the web repository of Wikipedia (Wiktionary[7]. The selection of these verb groups occurred because in these categories, the conjugation of most verbs from the third-

[6] For the French sentences, see Table 3.1

[7] https://en.wiktionary.org/wiki/Category:French_second_group_verbs

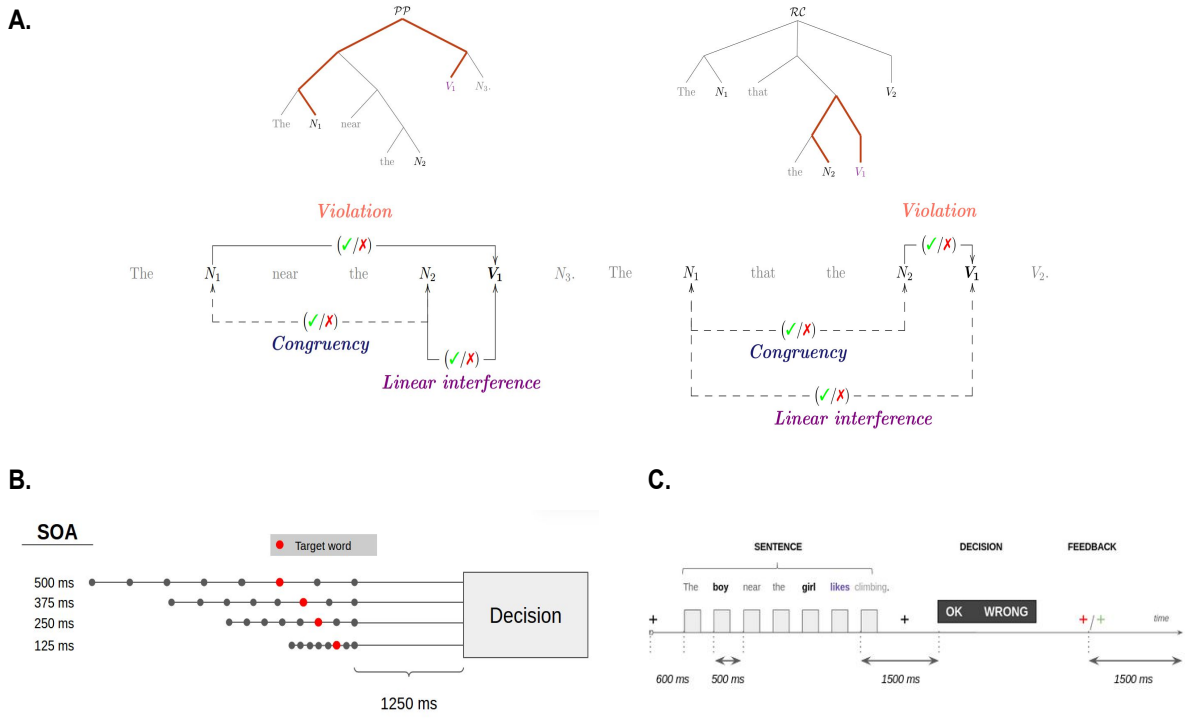


Figure 3.1: **Experimental design and paradigm**. To disentangle two possible types of processing during sentence comprehension, the experimental design contrast: (i) a structural dependency between a target verb and a noun, which defines the grammatical agreement (ii) a linear (word order) interaction between the target verb and another noun, which either facilitates or interferes with verb processing. **(A.) Tree representations of the two sentence constructions and illustration of the main effects of the design:** Violation effect (orange), which depends on the structural dependency between the subject and the verb (colored path in the tree representation). Interference effect (magenta), which refers to the (mis)match between the target verb and a linearly intervening noun. Congruency effect, which refers to the (mis)match between the two nouns; **(B.) M/EEG Experimental Paradigm:** Subjects performed a forced-choice, violation detection task in French, and in a setting with four different Stimulus Onset Asynchronies (SOAs) of $125ms, 250ms, 375ms$ and $500ms$. **(C.) Intracranial Experimental Paradigm:** Patients performed an RSVP, forced-choice, violation detection task in English, in a setting with a single SOA of $500ms$.

singular to the third-plural tense, requires a change of a multiple of letters, compared to the simple *"ent"* suffix addition that the first group requires.[8] After the corpus creation, a machine-learning, open-source conjucator[9] was applied to automatically conjugate all the words of the corpus. Next, verbs were filtered based on their log-unigram frequency[10], where infrequent words[11] were excluded from the lexicon. Additionally, words were selected such that their Levenshtein[12] distance was maximal, when conjugated from the singular to the plural form. A length criterion was also applied, such that only words between three and nine letters were kept in the lexicon. Finally, the lexicon and the constructed sentences were evaluated by native French speakers.

English Stimuli

For each construction, we presented 40 trials per run, for a total of 4 runs. Each run contained 90 trials (40 PP-Number, 40 ObjRC-Number plus 10 fillers. Each patient was presented with the same set of sentences. The English stimuli were also controlled for low-level features such as length and unigram frequencies. Whereas in our previous work (Chapter 4, we only used the definite English article, in this design, we also included demonstrative articles (this/these), in an attempt to maximize the contrast between the singular and the plural form.

M/EEG recordings

Recording took place in the NeuroSpin MEG center. 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, recording was recorded with a 60 and 64 channel MEG compatible Neuromag EEG cap, respectively. The brain signals were acquired at a sampling rate of 1000 Hz 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.

Intracranial recordings

Data were acquired from stereotactically placed depth electrodes (sEEGs) implanted for clinical purposes of seizure localization of pharmaco-resistant epilepsy. sEEG probes (PMT Corporation) were 0.8 mm in diameter, had 8–16 contacts and were implanted using a Robotic Surgical Assistant (ROSA; Medtech). Each contact was a platinum-iridium cylinder, 2.0 mm in length with a centre-to-centre separation of 3.5–4.43 mm. Each participant had multiple probes implanted. Intra-cranial data were collected using the NeuroPort recording system (Blackrock Microsystems), digitized at 2 kHz.

[8] For example, the verb *manger* (to eat) belongs to the first group and is conjugated as follows, from the third-singular to the third-plural tense: mange→mange**nt**. In contrast, the verb *convaincre* (to convince) that belongs in the third group has the following conjugation: convainc→convain**quent**.

[9] https://github.com/SekouD/mlconjug

[10] http://www.lexique.org/

[11] Unigram frequency less than three times the median unigram frequency

[12] A metric used in information theory, that describes the difference of two strings. For a definition and alternatives, see [Yujian and Bo, 2007].

Preprocessing

M/EEG

Bad sensors per sensor-type (magnetometers, gradiometers, eeg) were automatically detected at the run level based on a variance criterion. Channels for 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 artefacts were removed at the run level, using signal-space projection (SSP) implemented with MNE Python [Gramfort et al., 2013], [Jas et al., 2018]. To compensate for head movement and reduce non-biological noise, the MEG data were Maxwell-filtered [Taulu et al., 2004] using the implementation of Maxwell filtering in MNE Python. The bad EEG sensors were interpolated using the spherical spline method [Perrin et al., 1989] 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 45 Hz using a linear-phase FIR filter (hamming) with delay compensation, implemented in MNE-python version 0.16 [Gramfort et al., 2013]. Finally, the continuous time-series were segmented into 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. EEG data were re-referenced to the common average reference.

sEEG

Channels were visually inspected for line noise, artifacts and epileptic activity. The data were then bandpass filtered between 0.4 and 45 Hz using a linear-phase FIR filter (hamming) with delay compensation, implemented in MNE-python version 0.16 [Gramfort et al., 2013]. Finally, the continuous time-series were segmented into epochs of interest (first word onset to panel onset). Data were re-referenced to the common average reference of the non-rejected channels.

Decoding Analysis

We used a temporally defined decoding approach to classify neural activity from two conditions at the trial level [King and Dehaene, 2014], [Dehaene and King, 2016]. These analyses were implemented in MNE-python version 0.16 [Gramfort et al., 2013]. Prior to model fitting, the data were standardized using the Scikit-Learn package [Pedregosa et al., 2011]. We used a linear classifier (logistic-regression) with default Scikit-Learn parameters. The evaluation metric was the Area Under the Curve (AUC). Prior to decoding, the data were smoothed with a 100ms Gaussian kernel window. To prevent overfitting, we used a 5-fold stratified cross-validation procedure.

Statistical Analysis

The reported statistics correspond to group-level analyses and were performed using the Statsmodels package in Python3 and in MNE-python version 0.16 [Gramfort et al., 2013]. The statistical significance of the decoding performance over time was evaluated and corrected for multiple comparisons using a cluster-based permutation approach [Maris and Oostenveld, 2007], using a total of 1000 permutations. In this approach, a cluster is defined as adjoint time points that exceed a threshold of significance. The significance threshold (alpha level) for all analyses was set to 0.05.

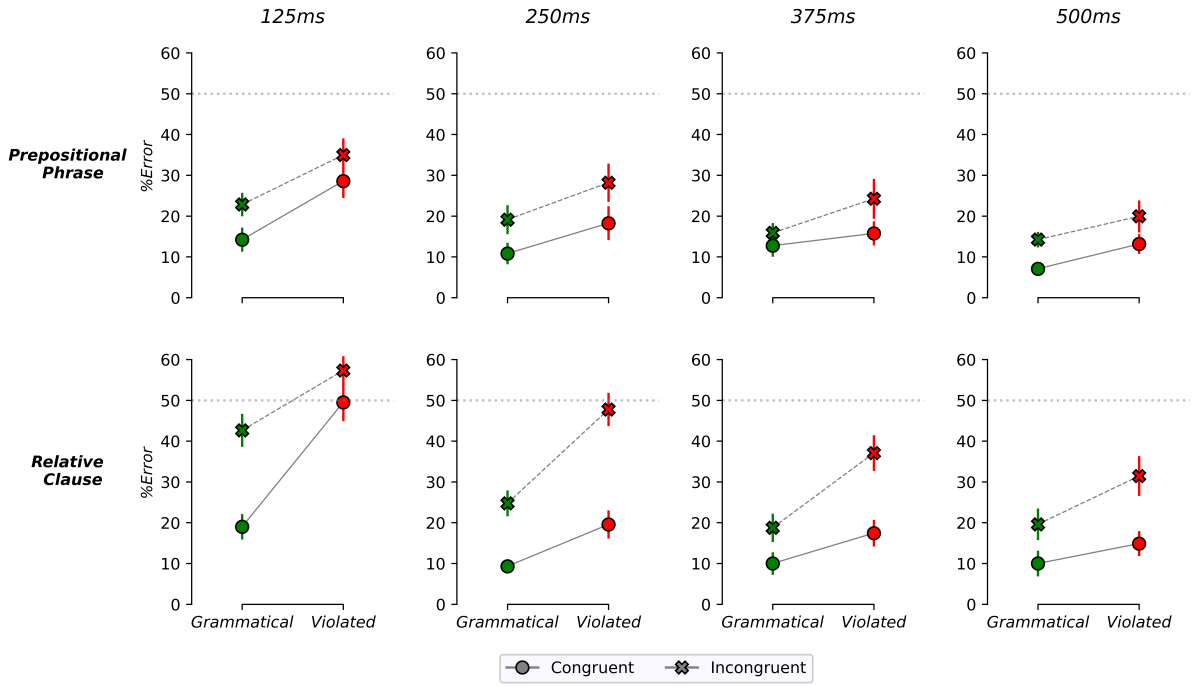


Figure 3.2: **Interaction plots for the Violation and Congruency effects** ($N = 20$), **M/EEG analysis.** The main effects for Violation and Congruency are systematically significant for the ObjRC-Number condition. This is not the case for the PP-Number condition, where no effect reaches significance for the 375ms SOA. The error bars indicate the standard error of the mean (SEM) calculated across participants.

	SOA	125ms		250ms		375ms		500ms	
		Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
PP-Number	Violation	**13.7**	**<e-03**	**4.6**	**<e-01**	n.s	n.s	**5**	**<e-01**
	Congruency	**4.4**	**<e-01**	**5.6**	**<e-01**	n.s	n.s	**7**	**<e-02**
	Interaction	n.s	n.s	n.s	n.s	n.s	n.s	n.s	n.s
ObjRC-Number	Violation	**34.2**	**<e-06**	**26.5**	**<e-05**	**13.3**	**<e-03**	**4.48**	**<e-01**
	Congruency	**16.6**	**<e-03**	**45.6**	**<e-08**	**16.2**	**<e-03**	**4.7**	**<e-01**
	Interaction	**4.2**	**<e-01**	n.s	n.s	n.s	n.s	n.s	n.s

Table 3.2: Eight one-way between subjects ANOVAs were conducted to compare the effects of congruency, violation, and their interaction (linear interference) on the error-rate.

3.3 Results

Following our previous work (Chapter 4, we define and analyze three main factors. Figure 4.1 provides a visual illustration of the main effects. The *Violation* effect controls the grammatical configuration of the sentence and can be attributed to structural operations, according to standard linguistic analysis. Notably, in our two linguistic modifiers, the range of the structural effect is not symmetric. In the case of the PP-modifier, this effect corresponds to a long-range dependency, whereas for the ObjRC, the effect can be realized as a transition between two adjacent words. Importantly, in this configuration, the transition probability between the adjacent words corresponds to a syntactically illicit template (Figure 4.1-Panel A). Additionally, our design allows for the definition of two additional effects. The *Congruency* effect corresponds to a relationship defined based on the inflectional covariation of the head noun and the second noun (hereafter *attractor*. The *Linear Interference* is defined as a dependency realized between the attractor and the verb. In the prepositional phrase construction, where the attractor is embedded between the head and the target verb, the linear interference effect coincides with a bigram transition between the attractor and the verb.

Behavioral Results

We first sought to identify the modulation of the error-rate by the main factors of our design. Thereefore, unlike Tanner et al. Figure 3.2 shows the interaction plots of the *Violation* and *Congruency* factors. Table 3.2 summarizes the results of a one-way, between subjects ANOVA analysis per cell of Figure 3.2. The *Violation* effect was significant across all SOAs for the ObjRC construction, but this was not the case for the PP, where the $375ms$ SOA did not lead to a significant violation effect. The effect was driven by the agrammatical sentences. Participants made more errors in detecting an existing violation, than confirming that a sentence was indeed grammatical. This phenomenon is known as a *grammatical illusion* [Wagers et al., 2009], as subjects experience the illusion of a grammatical sentence in the presence of violation, and thus make mistakes in assessing correctly the grammaticality of the sentence. [13] The *Violation* effect was not equally dominant across all four SOAs. This is especially evident in the ObjRC construction, where the faster the SOA, the stronger the effect was. The *Congruency* effect followed a similar profile to that of *Violation*. This effect was significant for all SOAs in the ObjRC construction and to all but $375ms$ for the PP. This effect was driven by the incongruency of the trials. Participants made more errors in sentences where the two nouns disagreed in grammatical number, compared to the ones where they did. [14] The interaction of the two factors is, leads by definition, to the third one. We only observed an effect of the attractor to the verb of the relevant grammatical agreement, in the ObjRC construction, and only for the fast SOA condition. Notably, this effect was driven by the fact that the congruent, agrammatical sentences were at chance level of performance. In contrast, in all other configurations the congruency of the violated sentence acted in a facilitatory way. In this configuration, the mere presence of a violation, inhibited severely the ability of the parser to perform the task. Interestingly, the parser, although driven close to chance level, managed to perform the task in the case of a grammatical incongruency. Notably, this pattern is not symmetric across linguistic constructions, as in the case of the PP construction, the performance of the subjects remained above chance. This illustrates the additional processing cost induced by the mere linguistic complexity of a structure, as it is known that relative clauses are generally harder to parse compared to prepositional phrases. We verified the structure effects across all

[13] For a detailed analysis, see table 3.5 and Figure 3.8.
[14] For a detailed analysis, see table 3.6 and Figure 3.10.

conditions, but the effects were more pronounced for the fast SOA. Figure 3.11 summarizes the results.

Additionally, we observed a clear effect of the SOA in the overall error-rate[15], with the fast SOA being more detrimental regardless of the linguistic construction. Notably, when analyzing only the filler trials, we see that the performance of the fast SOA is at chance level. We do not observe the same behavior when examining the canonical sentences (Figure 3.9a). This might be an indication that the subjects can already, at this level, develop task-resolving strategies. We return to this point on the discussion.

Finally, when it comes to the intracranial analysis, the task was initially difficult for the patients, with the first two patients performing below or almost at chance level (Figure 3.12). This led us to revise the paradigm and reduce the number of trials to its current version. In this chapter, we restrict the analysis to data originating from a single patient.

3.3.1 Neural correlates of the main factors, and a dependency on the correctness of the responses.

At the behavioral level, we observed a clear modulation of the behavioral responses by the two main factors of *Congruency* and *Violation*, across both structures. To identify the main effects in the neural data, we used standard decoding techniques: for each effect and each SOA, a linear, binary classifier was trained at each time point to separate trials from the two conditions, and then tested on unseen data in a cross-validation manner. Figures 3.3 and 3.4 present the results of our decoding analysis per linguistic structure and SOA, when taking all responses into account. Here, we parse the factorial design according to the main factors, and we seek the neural correlates of the main effects.

The prominence of the structural effect.

The *Violation* effect was detectable across both linguistic constructions, but unlike in the behavioral data, only for the SOAs of 375 and 500ms. The temporal profile of the effect was different both across SOA and construction, but it was also affected by the correctness of the responses. For the long-range dependency, the 375 SOA led to a significant effect 500ms after the onset of the target verb, that lasted for 400ms and reached its maximum AUC value at 860ms. The effect for the 500ms SOA became significant at 650ms after the onset of the verb, and lasted for 450ms, with a peak at 800ms. The structural effect for the ObjRC, was also, only evident for the slow SOAs (375 & 500ms), but the temporal profile of the decoding was different. For the 375ms, we observed a late emerging phenomenon with an onset of 900ms post verb presentation that lasted for 250ms. This was surprising, given the fact the *Violation* effect for this construction emerges from a dependency between adjacent words (Figure 4.1.) The 500ms SOA, led to an effect that surfaced at 550ms post-verb onset, with a duration of 180ms.

A neural correlate of *Congruency*

In contrast to our previous work, we managed to detect a direct, neural correlates of an effect stemming from the inflectional covariation of the two nouns. In particular, we detected a clear effect of *Congruency* in the case of the 125ms SOA and only for the PP-Number construction. This computation occurs, by definition, prior to the onset of the target-verb. It was therefore surprising to see such an effect emerging at such a late state, given that the AUC curve crossed

[15] For a different view on this, see Figure 3.9b.

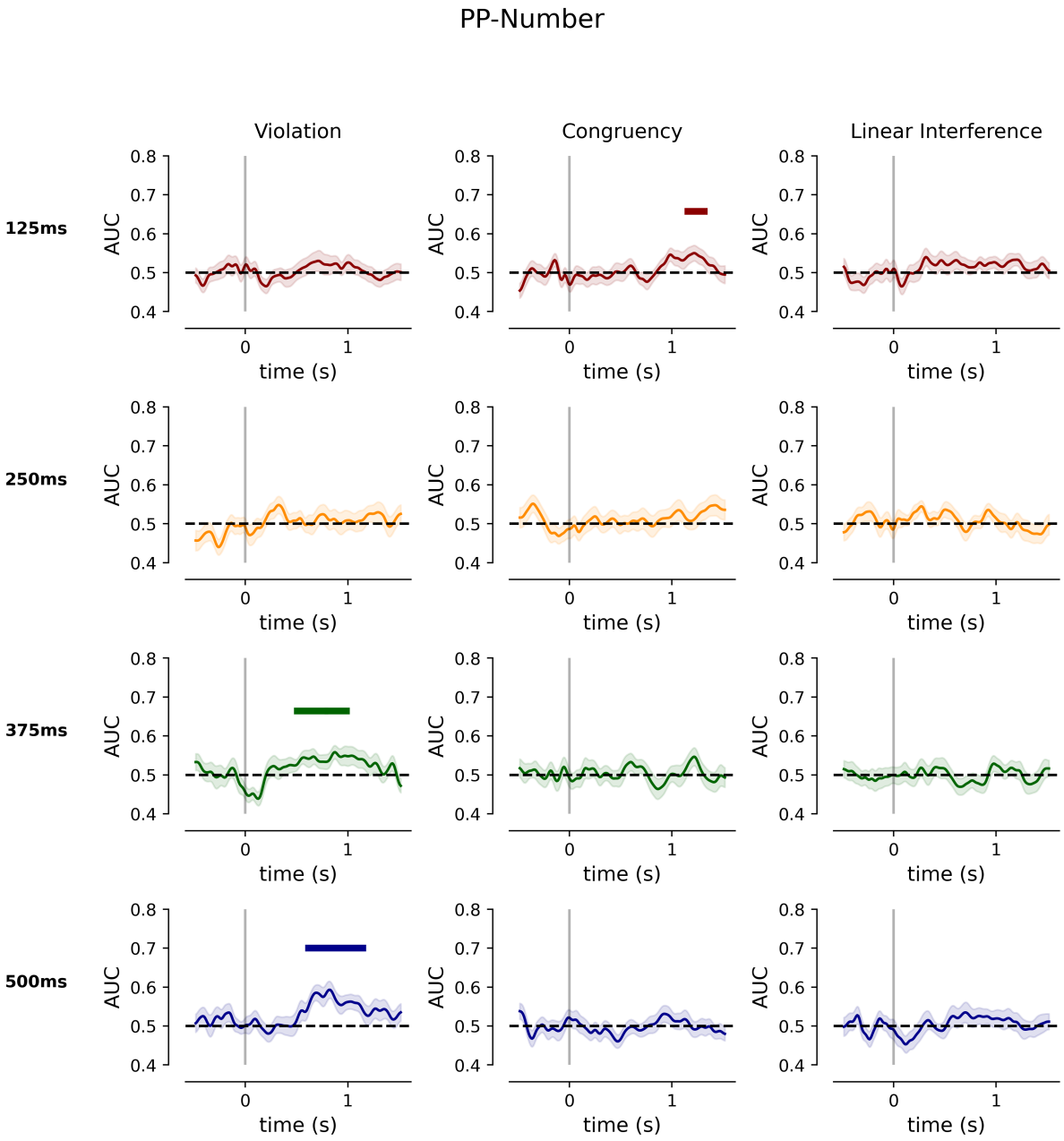


Figure 3.3: **Decoding of the main effects across SOAs, taking all responses into account, for the PP-Number construction.** Neural decoding of the three main effects using all sensor types (magnetometers, gradiometers and eeg). Time zero indicates the onset of the target verb. The period following the verb onset is the ISI to panel onset, common across all SOAs (Figure 4.1). A different, linear decoder was evaluated per time-point. The evaluation metric is the Area Under the Curve (AUC). The continues line above the classification plot indicates statistically significant time intervals ($p < 0.05$; corrected - spatio-temporal clustering permutation test). The main effect of *Violation* reaches significant classification accuracy for the slow SOAs (375&500ms) only. Unlike our previous work (Chapter: 4, we managed to detect direct neural correlates of a linear effect. The linear model reaches statistical significance for the *Congruency* effect, and only for the fastest SOA (125ns). The decoding of the second linear effect remained at chance level throughout the whole period of interest. Figure 3.13 summarizes the results when taking only the correct responses into account. For an analysis based on the false responses only, see Figure 3.15

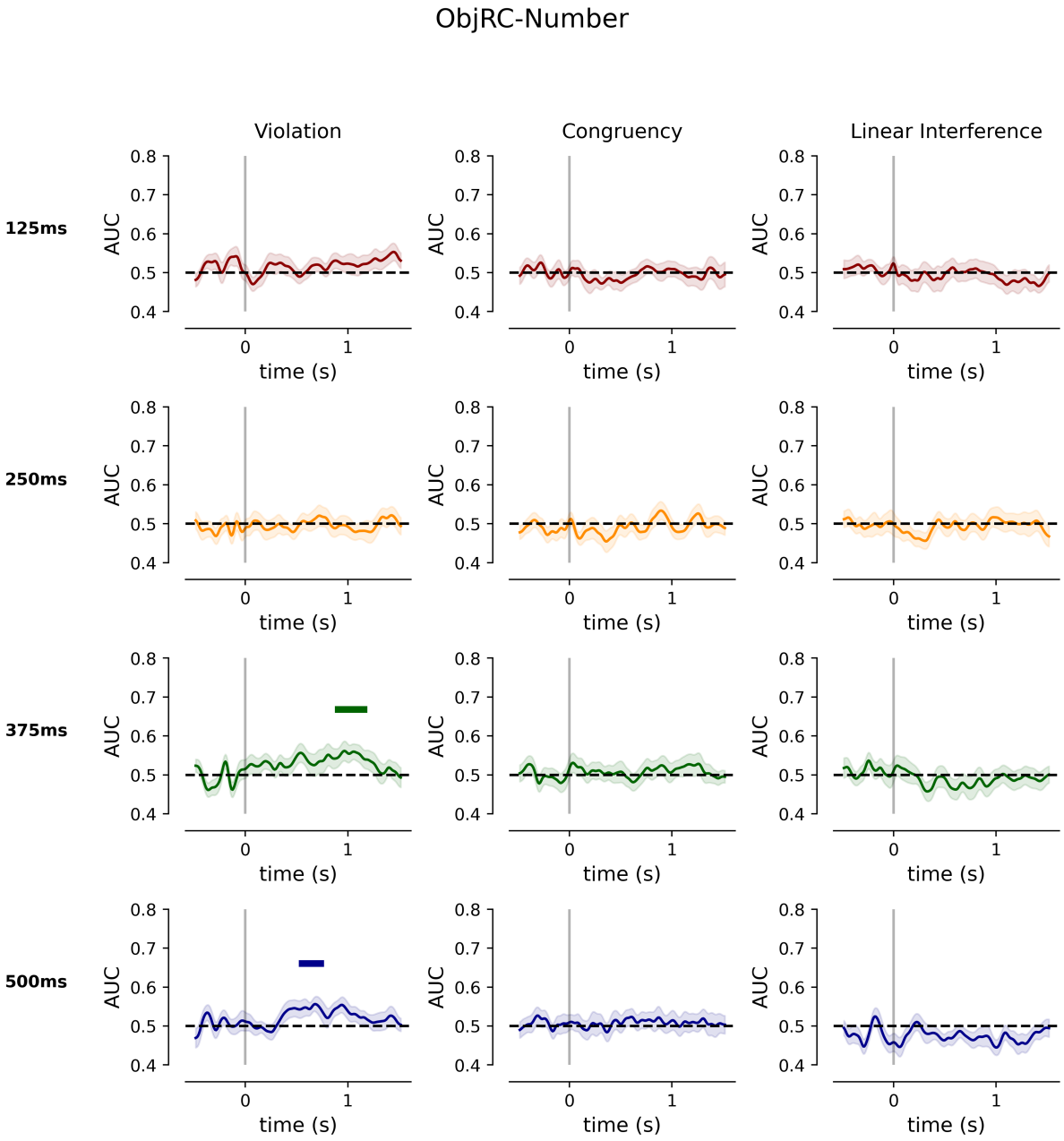


Figure 3.4: **Decoding of the main effects across SOAs, taking all responses into account, for the ObjRC-Number construction.** Neural decoding of the three main effects using all sensor types (magnetometers, gradiometers and eeg). Time zero indicates the onset of the target verb. The period following the verb onset is the ISI to panel onset, common across all SOAs (Figure 4.1). A different, linear decoder was evaluated per time-point. The evaluation metric is the Area Under the Curve (AUC). The continues line above the classification plot indicates statistically significant time intervals ($p < 0.05$; corrected—spatio-temporal clustering permutation test). The main effect of *Violation* reaches significant classification accuracy for the slow SOAs (375&500ms) only, similar to the PP-Number construction. Similar to our previous work (Chapter: 4, we did not manage to detect direct neural correlates of a non-structural factor. Figure 3.14 summarizes the results when taking only the correct responses into account. In this configuration, the *Congruency* effect becomes detectable in two SOAs (250&500ms), whereas the *Violation* effect is significant for three out of four SOAs (although the onset of the 375ms SOA appears to be strangely early). For an analysis based on the false responses only, see Figure 3.16

the significance threshold ($p < 0.05$ against chance - cluster-based permutation testing, [Maris and Oostenveld, 2007] at $1.15s$ after the onset of the verb. The effect had a duration $160ms$. The late onset of the effect might be suggestive of a re-analysis process, occurring after an initial assessment of the grammatical configuration has taken place.

The correctness of the response modulates the temporal profile of the effects.

When analyzing the main effects taking all responses into account, we observed the emergence of the structural factor for the relatively slow SOAs and a robust effect of *Congruency* in the fastest SOA. We then decided to further split the data by filtering based on the correctness of the subject responses, under the assumption that this filtering maximizes the selection of trials with the maximum cognitive engagement.

In agreement with what we reported in the all-responses analysis, we only managed to detect direct correlates of the *Violation* and *Congruency* effects for both constructions. Nevertheless, the onset of both effects regardless of the linguistic structure was earlier. In particular, for the PP-Number construction, this configuration led to the decodability of the structural effect for the $250ms$ SOA, something that was not possible when taking all responses into account. Remarkably, unlike all previous results, the effect for this SOA had a particularly late onset of significance, starting at $1.16s$ post-verb onset and lasting for $236ms$. For the SOA of $500ms$, the effect surfaced $150ms$ earlier compared to the all-responses analysis, starting at $510ms$ post-verb, with a duration of $600ms$. For the ObjRC construction, we were able to decode the *Violation* effect for all but the $125ms$ SOA. Notably, the onset of the effect for the $250ms$ SOA was later compared to the slower SOAs, as it became significant at $676ms$ post-target and lasted for $150ms$. The $375ms$ SOA showed a peculiar profile. The onset for this effect was particularly fast, surfacing exactly at the onset of the verb, with a second peak emerging much later at $700ms$. The two peaks had a comparable duration (356 and, $464ms$ respectively).

The unusual onset of the first wave is puzzling. Nevertheless, the fact that the *Violation* effect is decodable on this SOA in all response-filtering configurations (correct, false & all responses), speaks to the robustness and the psychological reality of this effect in this SOA.

Figures 3.13 & 3.14 summarize the main factor effects for this configuration.

A consistent, late profile of *Congruency*.

The *Congruency* factor was not decodable for the PP-Number construction when filtering for correct responses. Nevertheless, the effect had a clear and stable profile in the ObjRC-Number construction, for the $250ms$ and $500ms$ SOAs. Furthermore, the timing of the effect was consistent across both SOAs. In the $250ms$ case, the effect surfaced at $766ms$ after the onset of the verb and lasted for $180ms$. In the $500ms$ case, the profile was similar, with an emergence at $800ms$ and a duration of $250ms$. The timing of these effects is consistent with that of the PP-Number condition, when no response-filtering is applied (onset of $900ms$ and duration of $250ms$). In all cases, the *Congruency* effect follows the onset of the *Violation* effect. Remarkably, in our previous work, we could not decode this effect in the same settings (ObjRC-Number, $500ms$ SOA).

The *Violation* effect for the fast SOA only emerges in the false responses.

The ObjRC-Number construction was harder for the subjects compared to the prepositional phrase across all SOAs (Figure 3.11). In the case of the fast SOA, the subjects performed close to chance level (Figure 3.2). We thus searched for the neural correlates of the *Violation* effect in this configuration and selecting only on the trials in which the subjects gave a false response, as

we were unable to decode it in any other setting. Indeed, in this case, the *Violation* effect had a clear decoding profile, emerging at 500ms after the verb onset and lasting for 260ms. Figures 3.15 & 3.16 summarize the decoding analysis for this trial selection. Table 3.7 summarizes the timing of the first-order analysis across all trial and SOA selection.

The *Linear Interference* factor did not emerge at the neural level

Our main analysis sought the neural correlates of the main factors (Figure 4.1). At the behavioral level, the effects of *Violation* and *Congruency* were significant across both constructions and SOAs (apart from the 375ms SOA in the PP construction). The decoding results of this first-order analysis were in line with the behavioral evidence. Namely, by examining both constructions and SOAs, as well as the nature of the subject responses, we were only able to identify neural signatures of these two effects. No configuration led to the decodability of the main effect of *Linear Interference*. Simply put, in agreement with our previous work, we could not trace neural correlates of operations attributed to transition probabilities between linearly but not structurally adjacent words.

Modulation of the structural effect by the attractor interference.

In an attempt to boost the SNR of our analysis, we employed a second-order approach that builds on the dominance of the structural effect. In this approach, a linear model is trained on classifying *Violation* per SOA, but then, at test time, the pool of trials is separated based on the remaining two factors. For example, a classifier can be trained in classifying trials based on whether they contain a violation for all sentences of the PP-Number construction and the 500ms SOA, and then at test time, perform the same task only the incongruent trials. This would correspond to the trials pooled from rows 2vs3-Table 3.1. This approach allows us to, indirectly, investigate the modulation of the structural effect by the congruency factor. The same approach was followed for the factor of *Linear Interference*.

When performing the second-order analysis taking all responses into account, we observed a statistically significant modulation of *Violation*, by *Congruency* only in the case of the ObjRC-Number construction, and only for the slow SOA. The difference between the congruent and incongruent trials became significant at 690ms after the onset of the target, and remained significant for 130ms, in alignment with the overall late profile of the factor observed in our main analysis. The significance of the difference was evaluated using cluster-based permutation analysis [Maris and Oostenveld, 2007], where the chance level was set to zero. No other effect was observed for this configuration.

We then continued to perform the same analysis, screening the trials based on the correctness of the responses. In this configuration, only the effect of *Linear Interference* emerged, notably with a different temporal profile across the two constructions. Figures 3.5 & 3.6 summarize the results of this analysis. The difference was significant only in the case of the 125ms SOA in the PP-Number construction, and only for the 500ms SOA, in the case of the ObjRC. Notwithstanding, the onset of the relative differences was different for the two constructions.

The motivation behind the utilization of the multiple SOAs, and in particular the very fast one, was rooted in the assumption that effects rooted in simple statistical regularities, such as transition probabilities described in [Dehaene et al., 2015], might be too fast and therefore undetectable with slow SOA designs. The *Linear Interference* effect, in the case of the PP-Number construction, coincides with a pure transition probabilities effect (Figure 4.1). In this construction, we detected a clear modulation by this effect, in the 125ms SOA. The effect had a very early profile, with the relative difference becoming significant 120ms after the onset of the target verb, and a duration of 180ms. In the case of the ObjRC-Number construction,

the effect does not co-occur with a pure transition phenomenon, at least at the bi-gram level, but is an effect emerging from a dependency between the attractor, and the target verb. In this construction, and only for the $375ms$ we observed a very clear and sustained difference, emerging at a late stage. Specifically, the difference became significant $990ms$ after the onset of the verb, with a duration of $400ms$.

Although in this analysis, we observed a modulation of the structural effect by both factors, it is worth noting that the effects were not consistent across SOAs and constructions.

Preliminary intracranial results.

Our intracranial analysis was limited to a single patient, as s/he was the only one who completed the whole experiment. The patient had a bilateral fronto-temporal coverage (Figure 3.7 - Panel A). Similar to our previous, M/EEG experiment (Chapter : 4, the main effect of *Violation*, was the only effect for which the linear model reached high classification accuracy. Figure 3.7 presents the Generalization Across Time (GAT) matrices [King and Dehaene, 2014, Dehaene and King, 2016] for the main effects, across both constructions. The diagonal of the GAT matrix corresponds to the AUC curves reported throughout the results section.

The profile of the GAT matrix allows us to make an inference on the computational architecture that corresponds to the given to neural activity [King and Dehaene, 2014]. The diagonal profile of the *Violation* effect, for the ObjRC construction, points to heavily feed-forward architecture. This architecture necessitates that the likelihood and the prior information is systematically completed within each of the processing stage of the architecture. We can infer that, because the classifier does not generalize across time (should it have, the GAT profile would deviate from the diagonal activation). Simply put, this illustrates that different neuronal ensembles get activated sequentially and propagate the encoded information in a feed-forward manner. A generalization across time might have, for example, implied that a given neuronal population gets re-activated recursively under different firing distributions.

Similar to our main analysis, we sought to investigate the modulation of the structural effect by the presence of the attractor. We observed a clear modulation by the *Congruency* effect, but not from the *Linear Interference*. Importantly, these results should be interpreted cautiously, as sample size is minimal (N=1). Nevertheless, this preliminary analysis confirms the presence of the structural effect, the difficulty to directly decode a violation-of-transition effect, even in high SNR settings, and the modulation of the structural effect by the congruency factor.

3.4 Discussion

In this project, we sought to identify neural and behavioral correlates of two distinct mechanisms in language processing and comprehension. The first mechanism controls the grammatical configuration of the sentence and is attributed to purely structural operations [Vigliocco and Nicol, 1998, Rizzi, 2004]. Operations handled by this mechanism are commonly referred to as syntactic or, *hierarchical*. These operations are attributed to *recursion*, a uniquely human processing ability, hypothesized to be the core of human singularity [Chomsky, 2014a, Chomsky, 2009, Hauser et al., 2002]. This kind of processing is, often, axiomatic, in that many linguistic theories take it as a given [Uddén et al., 2020]. Nevertheless, this view is far from considered a consensus standard. A growing body of research supports the view that statistical regularities alone could suffice to explain linguistic phenomena usually attributed to hierarchical operations [Frank et al., 2012, Frank and Christiansen, 2018b]. In a behavioral study, Frank and Bod [Frank and Bod, 2011] demonstrated that reading times could be explained by probabilistic modelling alone, thus disregarding the necessity of a hierarchical processing. Operations that

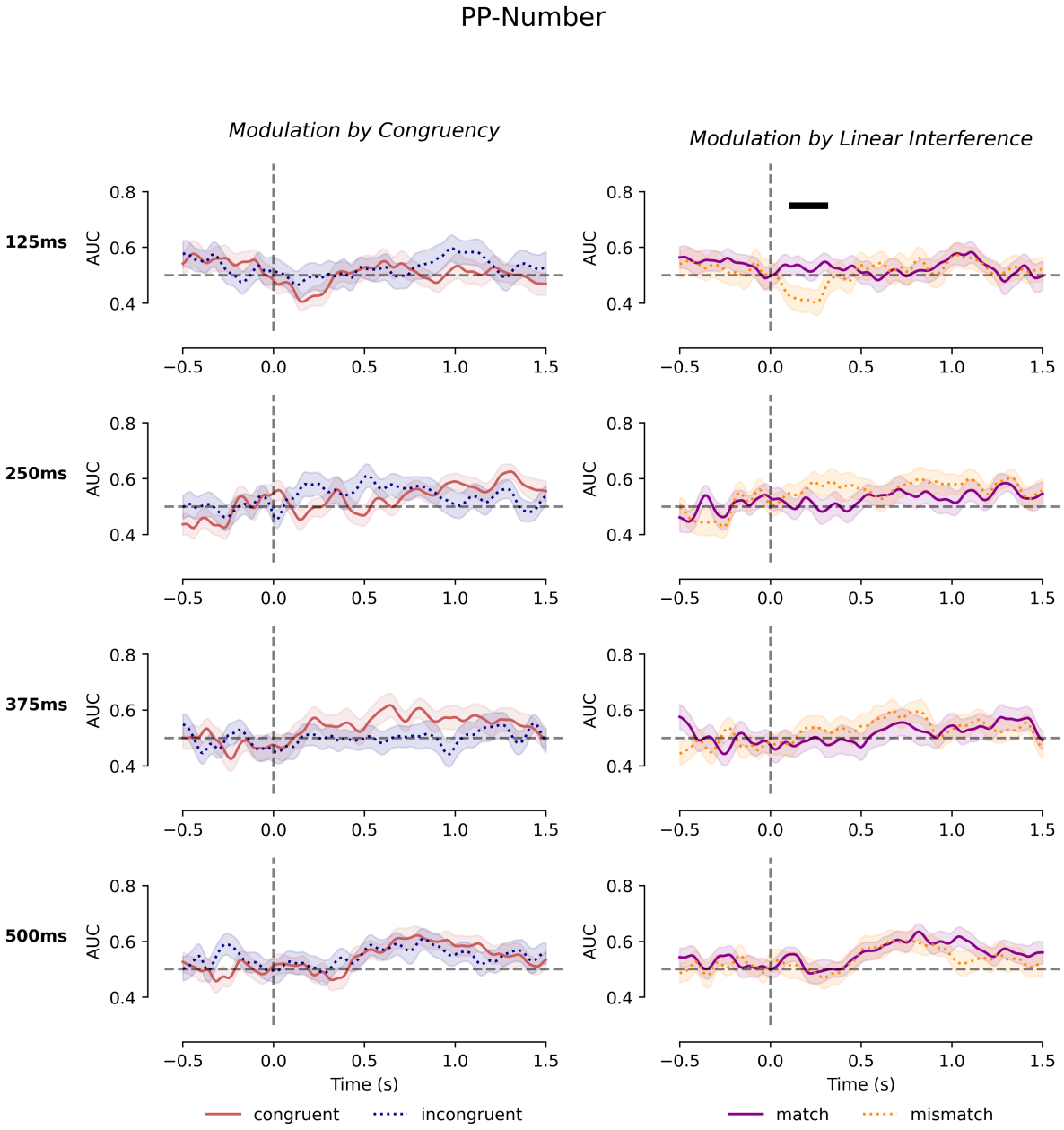


Figure 3.5: **Modulation of the structural effect by the attractor. A second order, decoding analysis across SOAs, taking only correct responses into account, for the PP-Number construction.** A linear model was trained on classifying neural data based on the presence or absence of a violation, and then, at test time, asked to classify a subset of this data in a cross-validated way. On the left column, the linear model was asked to classify violation separately for the *incongruent* trials (Table 3.1 rows 2 vs 3, dashed blue line), and the *congruent* trials (Table 3.1 rows 1 vs 4, continuous red line). This selection of the factorial provides an insight on how the congruency factor modulates the structural effect of violation. On the right column, the classifier was tested on a different subset. Instead of selecting trials based on their congruency, we separated the sentences based on whether the number of the non-head noun matched that of the target verb (Table 3.1 rows 1 vs 3, continuous magenta line) or whether there was a mismatch between the two (Table 3.1 rows 2 vs 4, continuous magenta line). This selection allows us to investigate the influence of the *Linear Interference* on the structural effect of *Violation*. Unlike our previous results (Chapter 4, we did not observe any modulation from the congruency factor on the long-range dependency. Nevertheless, we observed a statistically significant difference in the case of the extremely fast presentation condition, with an onset at 120ms after the verb presentation, and a duration of 180ms. Figure 3.17 presents the same analysis when taking all responses into account.

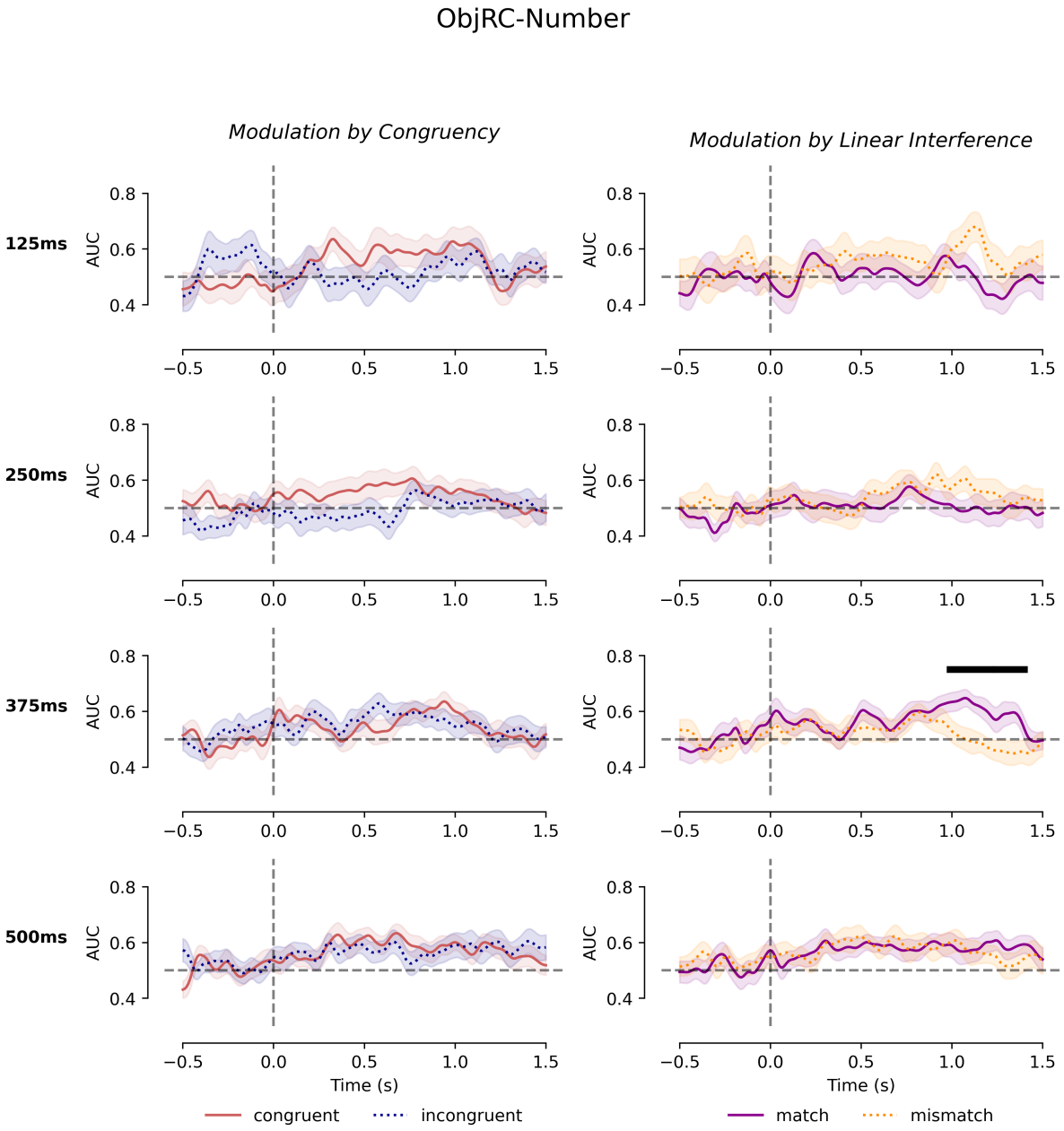


Figure 3.6: **Modulation of the structural effect by the attractor. A second order, decoding analysis across SOAs, taking only correct responses into account, for the ObjRC-Number construction.** On the left column, the linear model was asked to classify violation separately for the *incongruent* trials (Table 3.1 rows 5 vs 8, dashed blue line), and the *congruent* trials (Table 3.1 rows 6 vs 7, continuous red line). Unlike the PP-Number condition, and in agreement with our previous results, we observed a clear modulation of the structural effect by the congruency factor. This modulation stems from the fact that the decoding of the incongruent trials remained at chance level, whereas the congruent trials were easily decodable. The difference between the congruent and incongruent split was significant on the 375 and 500ms SOAs. On the right column, the classifier was tested on trials selected based on whether the number of the non-head noun matched that of the target verb (Table 3.1 rows 5 vs 7, continuous magenta line) or whether there was a mismatch between the two (Table 3.1 rows 6 vs 8, continuous magenta line). Similar to the PP-Number construction, we only observe a significant modulation from the effect of *Linear Interference*. Notably, in contrast to the long-range dependency, the effect here becomes significant at a very late stage (onset at 1s). Figure 3.18 presents the same analysis restricted to correct responses only, in which case, we observe a statistically significant difference in the 375ms SOA, similar to what we see in the PP-Number modifier.

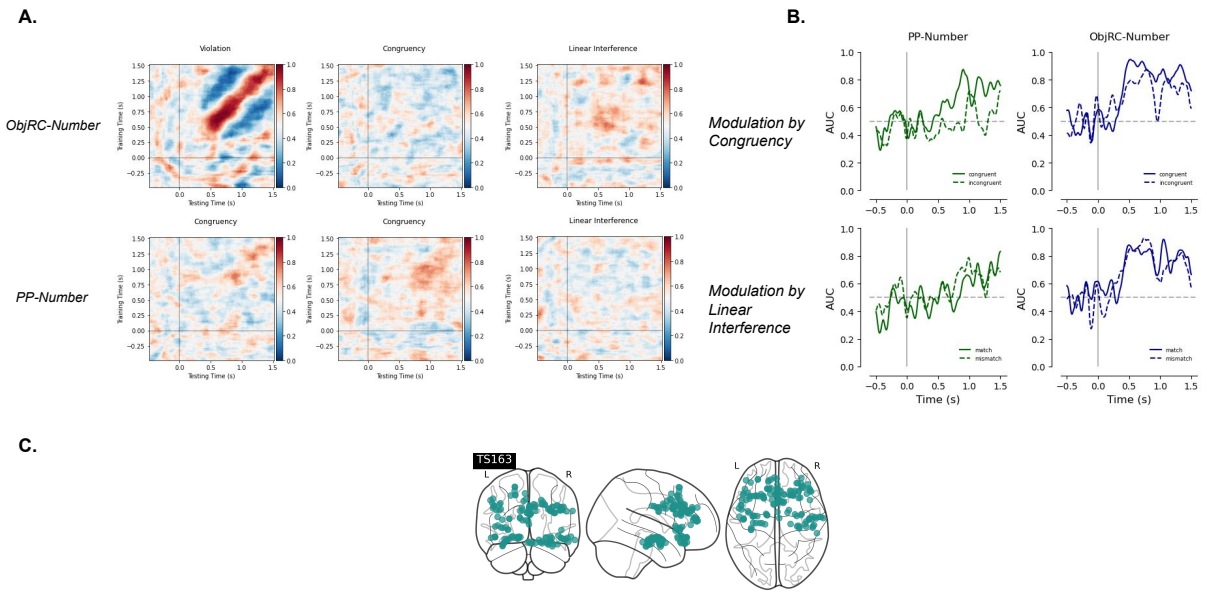


Figure 3.7: **(A.) Single patient, Generalization Across Time (GAT) matrices for the main effects and the two linguistic constructions. (C.) Modulation of the structural effect by congruency**. In this patient, the decodability of the structural effect for the ObjRC-Number condition reaches a phenomenal value, of almost 100% (AUC-leftmost GAT matrix). Despite this extremely high decodability of the *Violation* effect, we do not observe a similar profile for the other factors. Notably, in Panel (B.), and in agreement with our previous results, we observe a clear modulation by the *Congruency* factor. (C.) Electrode coverage of a single patient, consisting of 166 sEEG probes.

do not assume any underlying hierarchy and only occur at the sequence, or word-order level, realise the second mechanism. Operations attributed to this mechanism are often termed as *linear*. The implementations of this mechanism stem from an information-theory approach to language processing, and include metrics such as word-suprisal and entropy. For a review on probabilistic accounts of language processing, see [Armeni et al., 2017].

Syntactic operations have been traced in the human brain through a variety of studies. In a behavioral study, Shi et al. showed that toddlers can effectively understand hierarchical phrase structures, necessary to identify the grammatical configuration of two distinct linguistic structures [Shi et al., 2020]. In a recent study comparing humans and artificial language models, Coopmans et al. [Coopmans et al., 2021] showed that humans only interpret equivocal noun phrases such as "second blue ball" using a hierarchical parsing, in contrast to language models that require explicit hierarchical information during training. At the neural level, Pallier et al. [Pallier et al., 2011] presented word-sequences of varying lengths to adult volunteers during fMRI[16]. Their analysis led to an identification of a lef-lateralized, syntax specific network that includes the inferior frontal gyrus or "Broca's area" (IFG, Brodmann areas BA44 and 45) and the superior temporal sulcus (STS). In these areas, the activity increased monotonically as a function of the complexity of the phrase structure. Notably, the authors also included a "Jabberwocky" condition. In this condition, the sentences resembled normal sentences, but all the non-function words were replaced with meaningless tokens. In an ECOG[17] study, Nelson et al. [Nelson et al., 2017] analyzed high gamma activity from an RSVP experiment and identified sets of electrodes within the above-mentioned network, where the activity increased with each

[16] functional Magneric Resonance Imaging
[17] Electrocorticography

consecutive word, but abruptly diminished once a phrase could be completed. This result was interpreted as a correlate of the 'merge' operation, a core, syntax operation proposed by Noam Chomsky [Chomsky, 2014b], according to which, linguistic objects are unified into a common representation.

Despite the theoretical formulations and recent experimental evidence, the neural encoding of structural and linear operations remains an open challenge. Our project was designed to disentangle, operations attributed to both mechanisms. Our results draw a picture of language comprehension driven by structural operations robust to effects of low-level statistical regularities.

Nevertheless, evidence for a multiplex processing system that includes both structural and low-level transition probabilities, comes from two very recent neuroimaging studies. In an fMRI study, Cory Shain et al. [Shain et al., 2020] contrasted the fit of two models of surprisal[18] in fMRI, while subjects were listening to audiobooks. The two models consisted of a surface-based 5-gram model, and a hierarchical probabilistic context-free grammar (PCFG), essentially contrasting structural vs n-gram operations. Crucially, the authors found *significant* and *independent* results of both models in the Language Network (LANG: [Fedorenko et al., 2011, Fedorenko et al., 2012] but not the domain-general, fronto-parietal multiple demand network (MD: [Duncan, 2010]. Additionally, in a temporally refined, M/EEG study, Heilbron et al. [Heilbron et al., 2021] demonstrated that the brain employs prediction at the syntactic, semantic and phonemic level, thus providing evidence of probabilistic processing that occurs at a multiple of levels.

Our study differs from the above-mentioned in multiple aspects but most importantly, in both of these studies, the subjects were exposed to stimuli in more ecological settings (listening to audiobooks). The nature of our design was factorial and included a well-defined task. It might, thus be, that we failed to detect such effects because, due to the nature of our design, the subjects employed task-resolving strategies, which allowed them to circumvent the influence of such effects. Nevertheless, our results show a clear influence of the attractor both at the behavioral and the neural level. Should the subjects had solely relied on task-resolution strategies, we should not have observed any interference effects.

In a similar design that utilizes three different SOAs, Tanner and colleagues [Tanner et al., 2017] sought to investigate whether sentence level re-analysis processes, assumed to be reflected by the classical P600 ERP[19] component, and retrieval interference effects were distinguishable processes. First, Tanner and colleagues showed that interference (a term that corresponds to the incongruent trials of our design) reduced the amplitude of the component. Additionally, the authors reported a diminishing effect on the amplitude of the component, at faster SOAs. Importantly, Tanner et al. did not find a significant interaction between these two effects. This result was interpreted as evidence that retrieval interference and sentence re-analysis are distinct processes.

Even though we employed a multivariate analysis, contrary to the univariate, classical ERP approach of Tanner, our results are comparable and in partial agreement. The authors did not select trials based on the correctness of the responses, we would therefore compare our results as they pertain to this configuration. While Tanner et al. report a diminishing amplitude effect on faster SOAs, our decoding analysis reached statistical significance for the slow SOAs. The fact that we did not report a diminishing effect, but rather a no-effect, for the fast SOAs might be explained by the differences in the sample sizes between the two designs. While Tanner et

[18] Surprisal is an information-theoretic measure quantifying how unexpected the current word is, given the words that precede it. [Armeni et al., 2017]

[19] P600: Positivity occurring 600ms post verb onset, ERP: Event Related Potential

al. tested a total of 119 participants, our study was restricted to 20. The authors reported that the interference effect was diminishing on the amplitude of the P600. This analysis is equivalent to our second order analysis. In both the current and previous project (Chapter: 4, we observed a clear modulation of violation by congruency, with the incongruent trials leading to less decodability. In agreement with our main results, this modulation was no longer evident in faster SOAs. Therefore, unlike [Tanner et al., 2017], we observed an interaction between the SOA feature and the congruency effects. The late profile of the congruency effects point to the re-analysis stage of a two-step processing mechanism that does not affect the structural representation of the sentence, in agreement with behavioral results from Schlueter and colleagues [Schlueter et al., 2019]. The congruency effects surfaced consistently, after the structural effect of violation.

Furthermore, we only reported significant effects of congruency, that is, effects stemming from the factorial covariation of the inflectional morphology between the head-noun and the attractor, and not transition effects occurring between the attractor and the verb. These effects point to interference that stems from working memory operations. Behaviorally, we verified the phenomenon of grammatical asymmetry (the participants experienced attraction effects mostly in agrammatical sentences and not in grammatical). Amongst the families of attraction models, the cue-based model of attraction [McElree et al., 2003, Lewis and Vasishth, 2005, Van Dyke and Johns, 2012, Wagers et al., 2009, Badecker and Kuminiak, 2007, Dillon et al., 2013, Martin and McElree, 2008] is the model that can account for this phenomenon. This model, does not assume a continuous, global representation of the subject feature (unlike other accounts such as the representational models of attraction [Eberhard et al., 2005, Franck et al., 2002, Staub, 2009, Staub, 2010, Vigliocco and Nicol, 1998]. In contrast, attraction effects according to this model occur due to similarity-based interference at the retrieval stage. In agreement with previous work [Wagers et al., 2009], we observed congruency effects in the ObjRC construction, where the interference originates from a non-embedded attractor. It is important to note, that the cue-based retrieval model of attraction is the only model that can account for such effects.

3.5 Conclusion

We tackled an open question in language comprehension, and sought to isolate two discrete mechanisms of language processing and comprehension. Our results, corroborate the findings and conclusions of our previous work. Language processing appears to be driven by solely by structure-based computations and is robust to linear effects, such as transition-probabilities between non-structurally adjacent words.

Supplementary Material

3.6 Prototypical sentences used in the M/EEG & Intracranial Experiment

Construction	Template	Examples	Violation	Congruency
$PP - Number$	Det N_1 P det N_2 V_1 N_3	Le chef près du medecin craint le chien.	No	Yes
		Le chef près des medecins craint le chien.	No	No
		Le chef près des medecins craignent le chien.	Yes	No
		Le chef près du medecin craignent le chien.	Yes	Yes
$ObjRC - Number$	Det N_1 Adv det N_2 V_1 V_2	Le chef que le medecin craint part demain.	No	Yes
		Le chef que les medecins craignent part demain.	No	No
		Le chef que le medecin craignent part demain.	Yes	No
		Le chef que le medecin craignent part demain.	Yes	Yes
$PP - Filler$	Det N_1 P det N_2 V_1 N_3	Le chef près du doit craint le chien.	Yes	-
$ObjRC - Filler$	Det N_1 Adv det N_2 V_1 V_2	Le chef que le medecin craint pompier demain.	Yes	-

Table 3.3: Prototypical sentences used in the M/EEG experiment.

Construction	Template	Examples	Violation	Congruency
$PP - Number$	Det N_1 P det N_2 V_1 N_3	The doctor near the nurse likes climbing.	No	Yes
		The doctor near the nurses likes climbing.	No	No
		The doctor near the nurses like climbing.	Yes	No
		The doctor near the nurse like climbing.	Yes	Yes
$ObjRC - Number$	Det N_1 Adv det N_2 V_1 V_2	The doctor that the nurse fears leaves.	No	Yes
		The doctor that the nurses fear leaves.	No	No
		The doctor that the nurses fears leaves.	Yes	No
		The doctor that the nurse fear leaves.	Yes	Yes
$PP - Filler$	Det N_1 P det N_2 V_1 N_3	The doctor near the sneezes likes climbing.	Yes	-
$ObjRC - Filler$	Det N_1 Adv det N_2 V_1 V_2	The doctor that the nurse fears farmer.	Yes	-

Table 3.4: Prototypical sentences used in intracranial experiment.

3.7 Additional behavioral analysis.

SOA:		**125ms**	**250ms**	**375ms**	**500ms**
PP	Statistic	22.5	17.5	n.s	44.0
	p-value	2.19e-03	1.96e-03	n.s	2.62e-02
ObjRC	Statistic	1.0	6.0	27.5	37.0
	p-value	3.81e-06	2.35e-04	4.04e-03	1.17e-02

Table 3.5: A series of Wilcoxon test (paired samples) with Bonferroni correction that verify the *grammatical illusion* phenomenon in the M/EEG behavioral data.

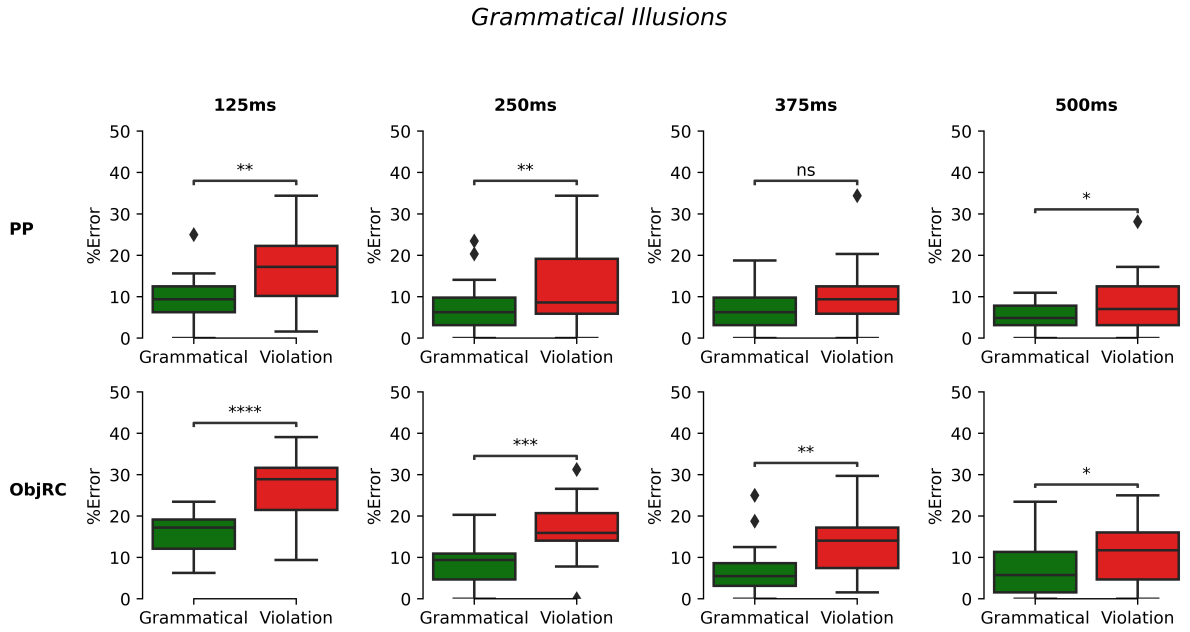
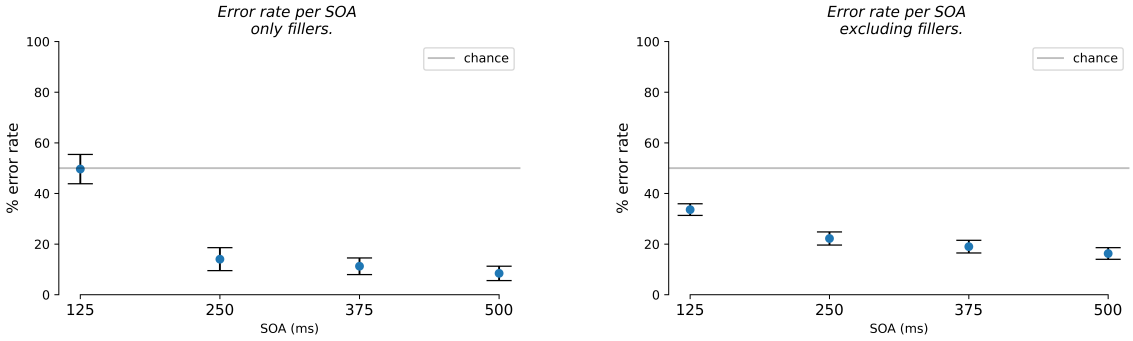


Figure 3.8: **Grammatical Illusion**. Subjects made more mistakes in detecting a violation compared to confirming that a sentence was indeed grammatical. This phenomenon is called *grammatical illusion*, as there appears to be an illusion of grammaticality in the violated sentences [Wagers et al., 2009]. The corresponding statistics are presented in table 3.5.



(a) label 1 (b) label 2

Figure 3.9: **Overall accuracy per SOA.** Subjects failed to perform the grammaticality judgement task on the filler trials and the case of the fast SOA, but not in the case of the normal sentences. These might be indicative of a strategy e on behalf of the participants, a strategy in which the encoding of the head noun and the corresponding verb is sufficient to do the task successfully, without reading the whole sentence.

SOA:		**125ms**	**250ms**	**375ms**	**500ms**
PP	Statistic	29.5	1.0	26.0	12.0
	p-value	5.06e-03	1.34e-04	1.81e-02	9.47e-04
ObjRC	Statistic	18.0	1.0	3.0	13.0
	p-value	1.23e-03	3.81e-06	1.50e-04	7.19e-04

Table 3.6: A series of Wilcoxon test (paired samples) with Bonferroni correction that verify that subjects made more errors in incongruent trials compared to the congruent ones.

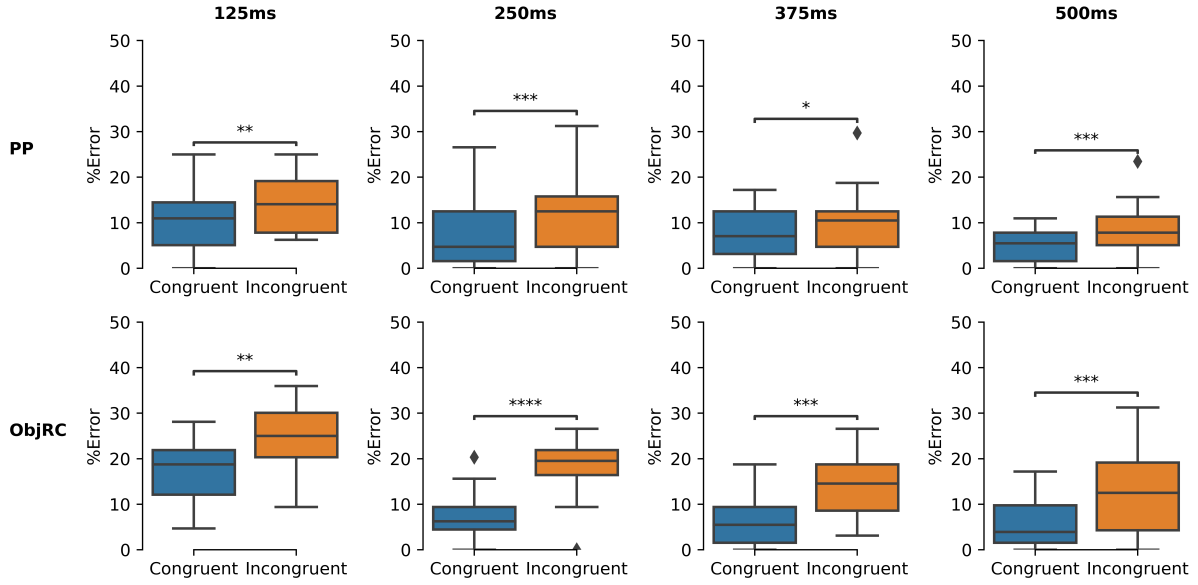


Figure 3.10: **Incongruent sentences led to more errors**. Subjects made more mistakes in performing the task when a sentence was incongruent compared to when it was congruent. The corresponding statistics are presented in table 3.6.

3.8 The correctness of the responses modulates the profile of the neural effects.

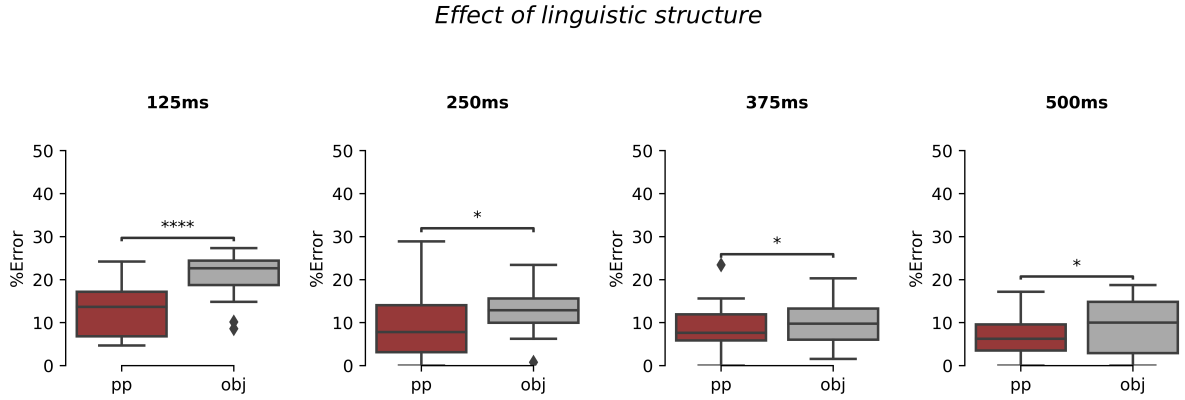


Figure 3.11: **Object relative clauses lead to more errors compared to prepositional phrases**. We verified the linguistic structure effect across all SOAs, with the effects being more prominent in the fast SOA.

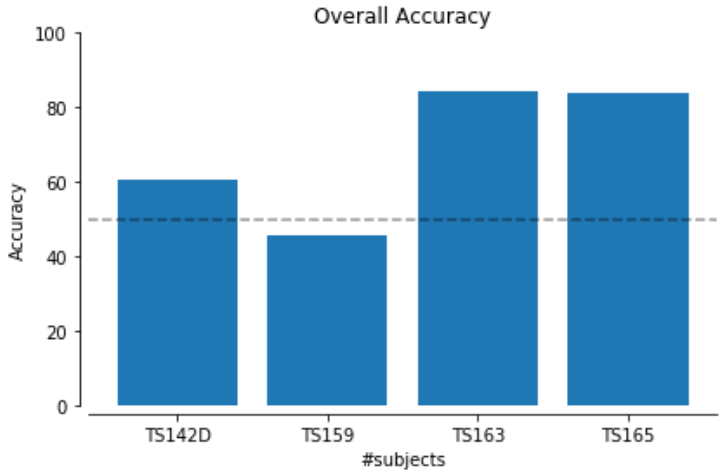


Figure 3.12: **Overall behavioral accuraccy for the patients**. The task was very hard for the patients. We, therefore, had to reduce the task duration in order to get meaningful responses. In this chapter, we present single-subject results originating from patient TS163.

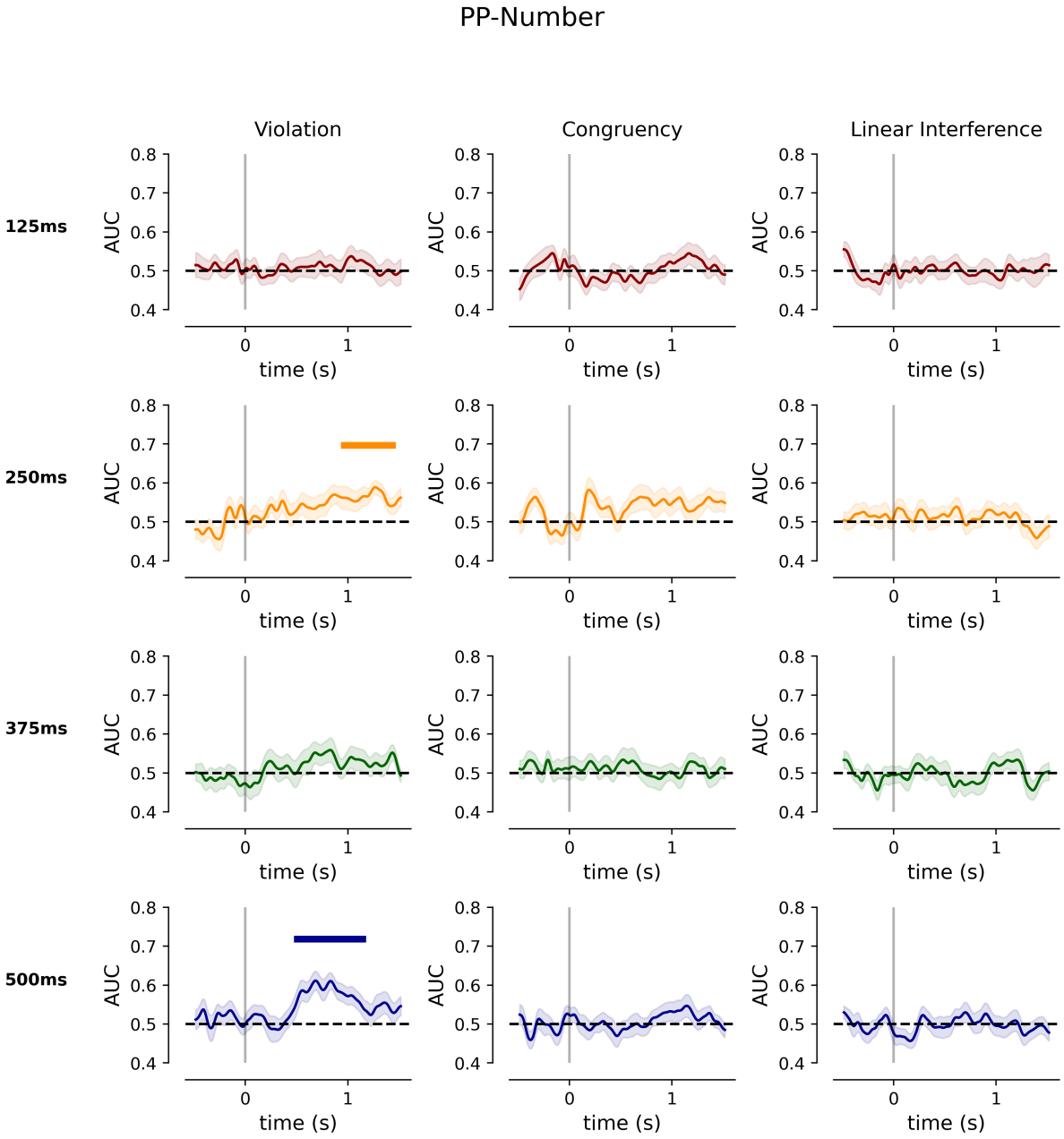


Figure 3.13: **Decoding of the main effects across SOAs, taking only correct responses into account, for the PP-Number construction.** Neural decoding of the three main effects using all sensor types (magnetometers, gradiometers and eeg). Time zero indicates the onset of the target verb. The period following the verb onset is the ISI to panel onset, common across all SOAs (Figure 4.1. A different, linear decoder was evaluated per time-point. The evaluation metric is the Area Under the Curve (AUC). The continues line above the classification plot indicates statistically significant time intervals ($p < 0.05$; corrected—spatio-temporal clustering permutation test). The main effect of *Violation* reaches significant classification accuracy for two out of four SOAs. Notably, the linear effect of congruency is no longer decodable. Additionally, the effect for the $375ms$ does not reach statistical significance, unlike when taking all responses into account (Figure 3.3). Interestingly, the $375ms$ decoding curve presents a late profile, compared to the $500ms$. For an analysis based on the false responses only, see Figure 3.15

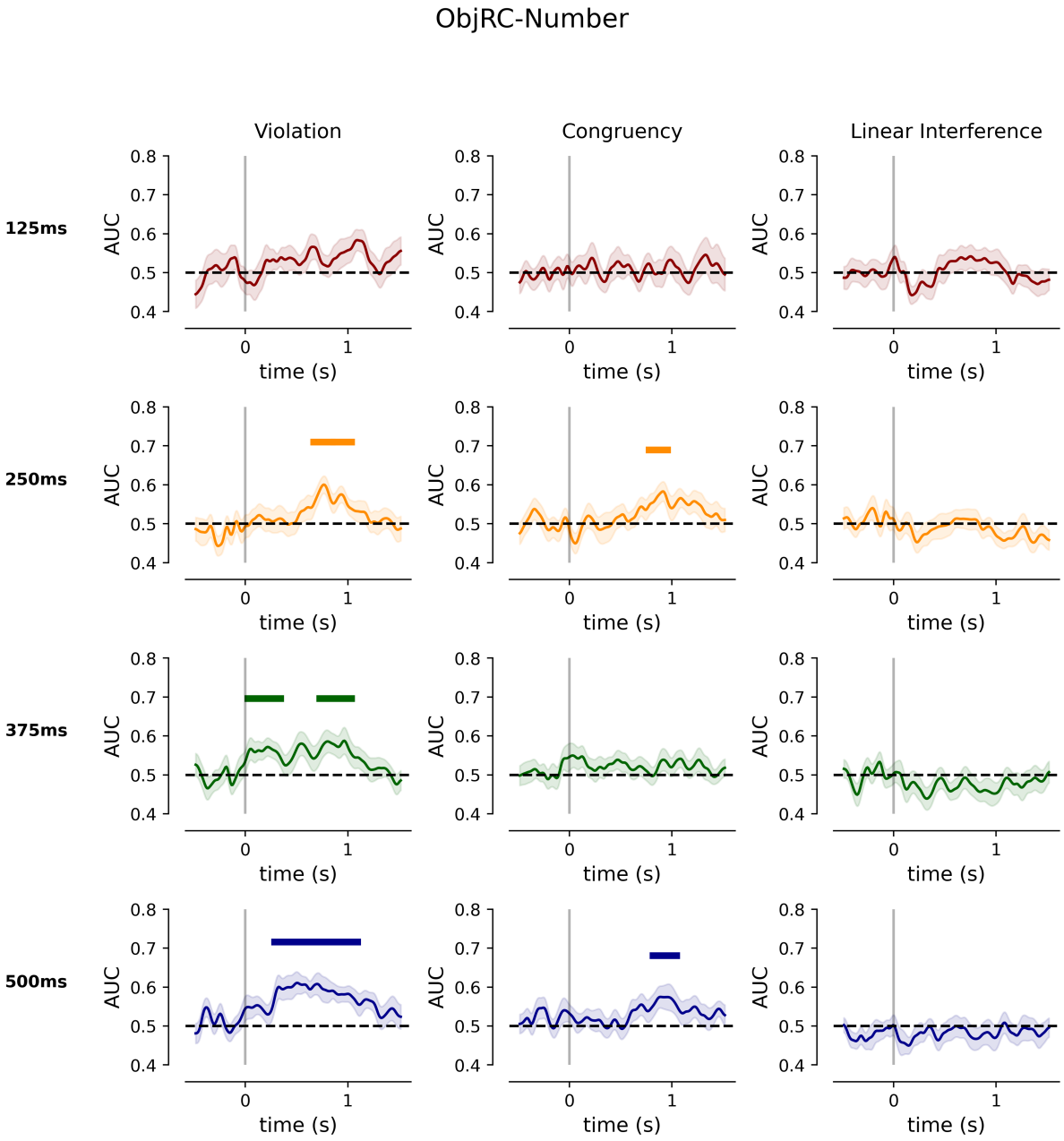


Figure 3.14: **Decoding of the main effects across SOAs, taking only correct responses into account, for the ObjRC-Number construction.** Neural decoding of the three main effects using all sensor types (magnetometers, gradiometers and eeg). Time zero indicates the onset of the target verb. The period following the verb onset is the ISI to panel onset, common across all SOAs (Figure 4.1). A different, linear decoder was evaluated per time-point. The evaluation metric is the Area Under the Curve (AUC). The continues line above the classification plot indicates statistically significant time intervals ($p < 0.05$; corrected—spatio-temporal clustering permutation test). The main effect of *Violation* reaches significant classification accuracy for all but the 125ms SOA. Importantly, the linear effect of congruency is decodable for the 250 and 500ms SOAs, with a late significant peak, similar to what observed for the 125ms SOA in Figure 3.3. For an analysis based on the false responses only, see Figure 3.15. For an analysis where no response-filtering is applied, see Figure 3.4.

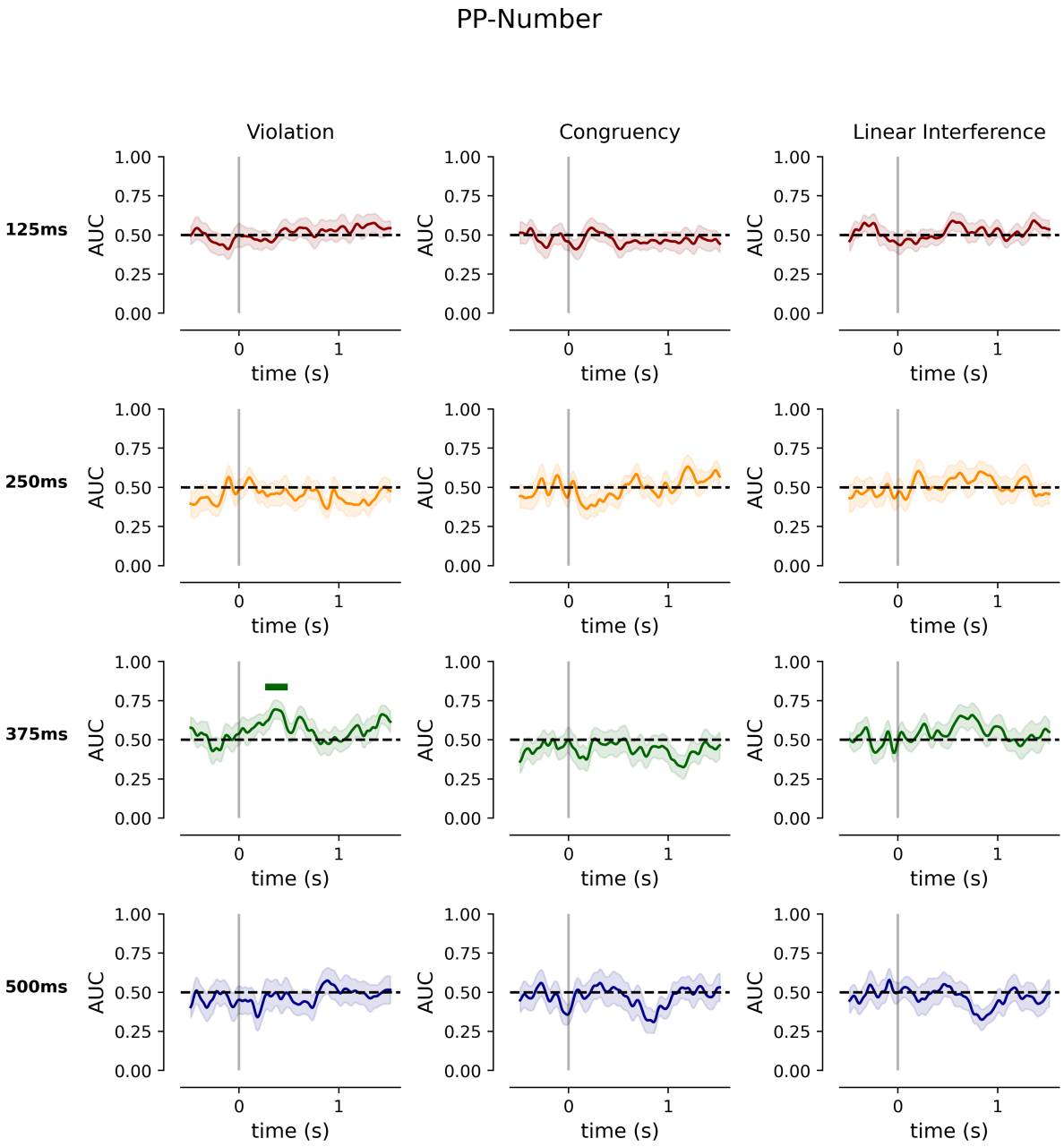


Figure 3.15: **Decoding of the main effects across SOAs, taking only false responses into account, for the PP-Number construction.** Neural decoding of the three main effects using all sensor types (magnetometers, gradiometers and eeg). Time zero indicates the onset of the target verb. The period following the verb onset is the ISI to panel onset, common across all SOAs (Figure 4.1. A different, linear decoder was evaluated per time-point. The evaluation metric is the Area Under the Curve (AUC). The continues line above the classification plot indicates statistically significant time intervals ($p < 0.05$; corrected—spatio-temporal clustering permutation test). The main effect of *Violation* reaches significant classification accuracy only for the 375ms SOA, which was not the case when examining only the correct responses for this construction (Figure 3.13. For an analysis based on all responses, see Figure 3.3

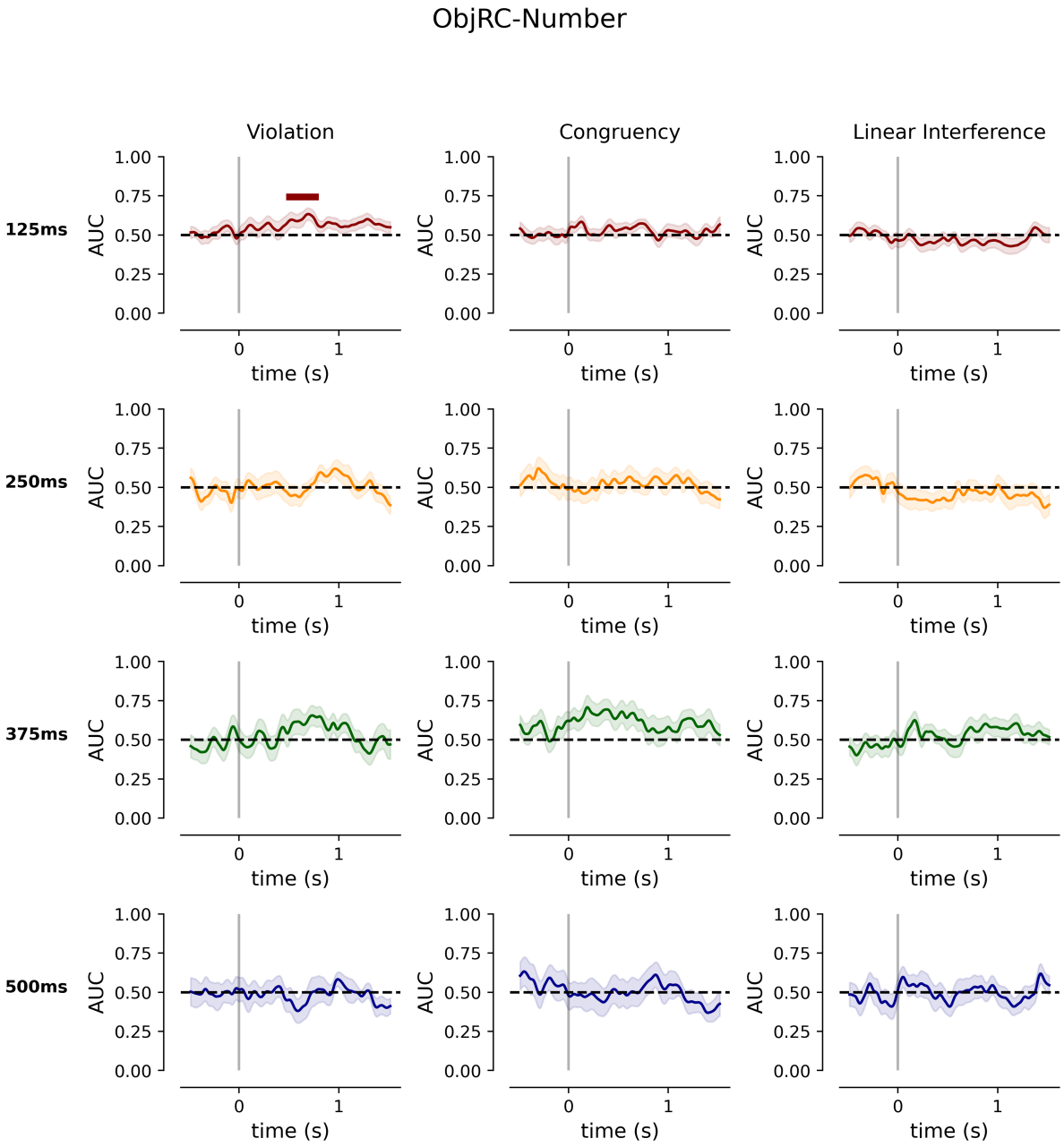


Figure 3.16: **Decoding of the main effects across SOAs, taking only false responses into account, for the ObjRC-Number construction.** Neural decoding of the three main effects using all sensor types (magnetometers, gradiometers and eeg). Time zero indicates the onset of the target verb. The period following the verb onset is the ISI to panel onset, common across all SOAs (Figure 4.1). A different, linear decoder was evaluated per time-point. The evaluation metric is the Area Under the Curve (AUC). The continues line above the classification plot indicates statistically significant time intervals ($p < 0.05$; corrected—spatio-temporal clustering permutation test). The only decodable effect is that of *Violation*, that reaches significant classification accuracy only for the 125ms SOA only. This is the only configuration that leads to the decodability of the *Violation* effect for this SOA. For an analysis based on all responses, see Figure 3.4

			PP			ObjRC		
Response Type	*Effect*	SOA	Onset time in (s)	Peak Time in (s)	Duration in (ms)	Onset time in (s)	Peak Time in (s)	Duration in (ms)
All Responses	Violation	125ms	-	-	-	-	-	-
		250ms	-	-	-	-	-	-
		375ms	0.5	0.86	400	0.9	0.97	256
		500ms	0.65	0.8	450	0.556	0.93	180
	Congruency	125ms	1.156	1.216	160	-	-	-
		250ms	-	-	-	-	-	-
		375ms	-	-	-	-	-	-
		500ms	-	-	-	-	-	-
	Linear Interference	125ms	-	-	-	-	-	-
		250ms	-	-	-	-	-	-
		375ms	-	-	-	-	-	-
		500ms	-	-	-	-	-	-
Correct Responses	Violation	125ms	-	-	-	-	-	-
		250ms	1.16	1.256	236	0.676	0.76	150
		375ms	-	-	-	0.05, 696	0.806	356, 464
		500ms	0.516	0.82	597.6	0.276	0.506	820
	Congruency	125ms	-	-	-	-	-	-
		250ms	-	-	-	0.766	0.92	180
		375ms	-	-	-	-	-	-
		500ms	-	-	-	0.806	0.87	250
	Linear Interference	125ms	-	-	-	-	-	-
		250ms	-	-	-	-	-	-
		375ms	-	-	-	-	-	-
		500ms	-	-	-	-	-	-
False Responses	Violation	125ms	-	-	-	0.506	0.696	260
		250ms	-	-	-	-	-	-
		375ms	0.296	0.436	160	-	-	-
		500ms	-	-	-	-	-	-
	Congruency	125ms	-	-	-	-	-	-
		250ms	-	-	-	-	-	-
		375ms	-	-	-	-	-	-
		500ms	-	-	-	-	-	-
	Linear Interference	125ms	-	-	-	-	-	-
		250ms	-	-	-	-	-	-
		375ms	-	-	-	-	-	-
		500ms	-	-	-	-	-	-

Table 3.7: **Timing information for the main effects and all response types.** In this comparative table, we gather timing information for time-intervals defined as statistically significant based on the evaluation of the decoding AUC curve against chance level. The statistical significance is corrected for multiple comparisons and calculated based on cluster-based permutation testing [Maris and Oostenveld, 2007].

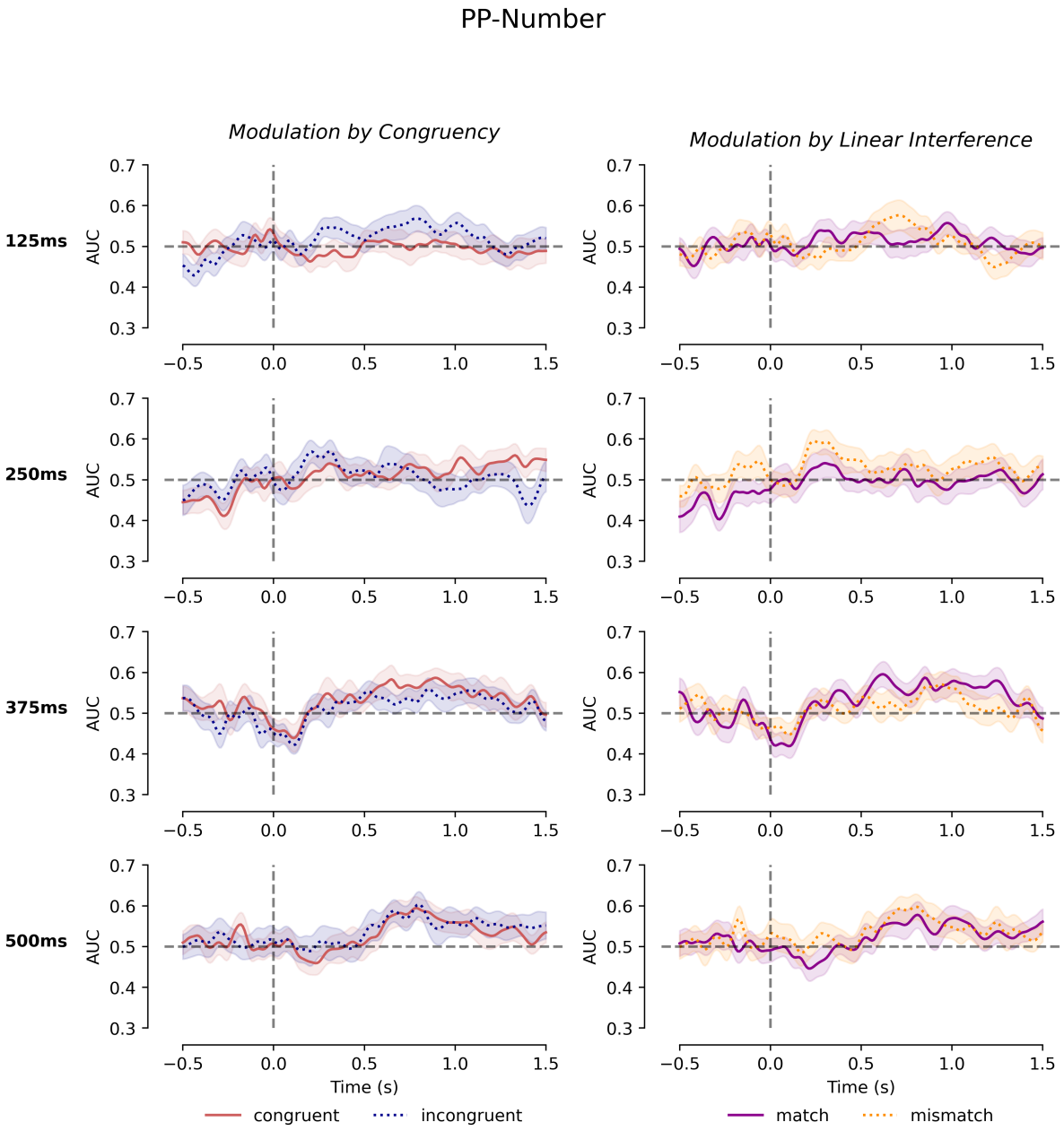


Figure 3.17: **Modulation of the structural effect by the attractor. A second order, decoding analysis across SOAs, taking all responses into account, for the PP-Number construction.** A linear model was trained on classifying neural data based on the presence or absence of a violation, and then, at test time, asked to classify a subset of this data in a cross-validated way. On the left column, the linear model was asked to classify violation separately for the *incongruent* trials (Table 3.1 rows 2 vs 3, dashed blue line), and the *congruent* trials (Table 3.1 rows 1 vs 4, continuous red line). This selection of the factorial provides an insight on how the congruency factor modulates the structural effect of violation. On the right column, the classifier was tested on a different subset. Instead of selecting trials based on their congruency, we separated the sentences based on whether the number of the non-head noun matched that of the target verb (Table 3.1 rows 1 vs 3, continuous magenta line) or whether there was a mismatch between the two (Table 3.1 rows 2 vs 4, continuous magenta line). This selection allows us to investigate the influence of the *Linear Interference* on the structural effect of *Violation*. Unlike our previous results (Chapter 4), we did not observe any modulation from the linear factors on the long-range dependency. Figure **??** presents the same analysis restricted to correct responses only, in which case, we observe a statistically significant difference in the 125ms SOA.

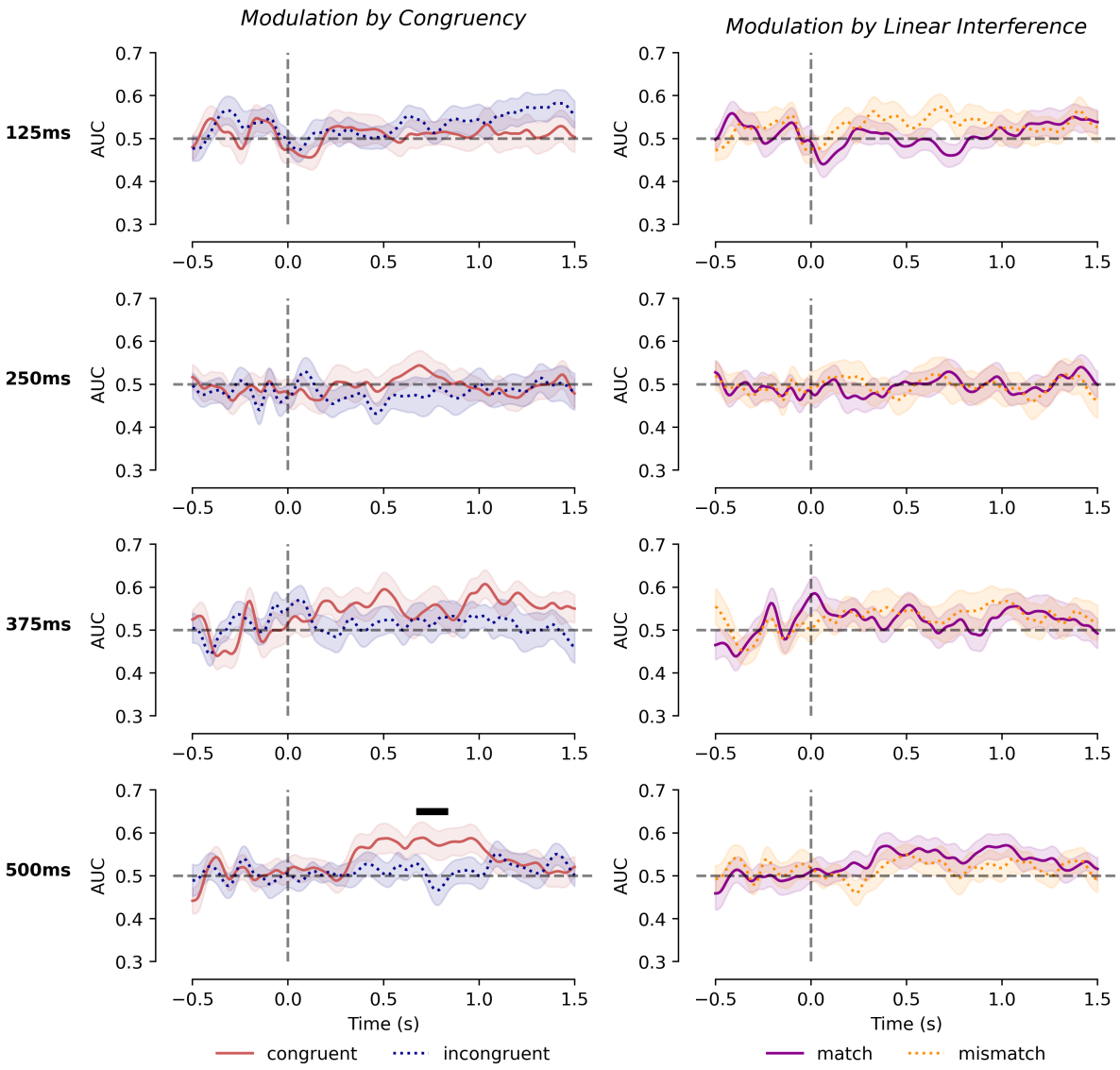


Figure 3.18: **Modulation of the structural effect by the attractor. A second order, decoding analysis across SOAs, taking all responses into account, for the ObjRC-Number construction.** On the left column, the linear model was asked to classify violation separately for the *incongruent* trials (Table 3.1 rows 5 vs 8, dashed blue line), and the *congruent* trials (Table 3.1 rows 6 vs 7, continuous red line). Unlike the PP-Number condition, and in agreement with our previous results, we observed a clear modulation of the structural effect by the congruency factor. This modulation stems from the fact that the decoding of the incongruent trials remained at chance level, whereas the congruent trials were easily decodable. The difference between the congruent and incongruent split was significant on the 500ms SOAs On the right column, the classifier was tested on trials selected based on whether the number of the non-head noun matched that of the target verb (Table 3.1 rows 5 vs 7, continuous magenta line) or whether there was a mismatch between the two (Table 3.1 rows 6 vs 8, continuous magenta line). There was no effect of *Linear Interference* in this configuration, nevertheless, we observed a very clear, late effect when taking correct responses only into account. Figure **??** presents these results.

3.9 Questionnaire & Responses

Did you find a way to complete the task without reading the entire sentence?

- 71.4% of the participants replied YES.
- 28.6% of the participants replied NO.

If you answered "YES" to the previous question, please explain your strategy here.

Original Replies

- Je me base sur un resentit et une impression lorsque c'est trop rapide.
- En fonction de si le verbe était au pluriel ou singulier (mais pas valable pour tout) en cherchant le sujet et les verbes.
- Identifier si les groupes nominaux étaient pluriels ou singuliers, et vérifier la correspondance avec le/les verbes. Pour cela, les determinants et les terminaisons des verbes suffisent.
- Reperer le singulier et le pluriel.
- Les types d'erreur étaient souvent les mêmes : mauvaise conjugaions de verbe ou bien mots incohérent au milieu de la phrase. En faisant attention à ça j'essayais d'obtenir la réponse quand la phrase défilait trop vite.
- L'ortographe.
- En regardant juste le nombre de sujets au pluriel et de verbes au pluriel.
- Je me concentrais sur le sujet de la phrase, qui était rapidement faussé ou validé par le verbe qui arrivait.
- Dès l'erreur arrivée, je stop la lecture de la phrase. Si pas d'erreur, lecture complète.
- En identifiant le sujet et l'accord du verbe et parfois dès l'apparition d'un non sens dans la phrase lire les deux noms et retenir si pluriel ou singulier puis voir si les verbes s'accordaient.
- J'essayais de reperer les singuliers et plurielles ainsi que les sujets pour voir si cela correspondait.
- Je faisais surtout attention à l'accord pluriel/singulier en me concentrant sur le sujet et le verbe à partir du moment où il y avaient des noms au pluriel, dès qu'un verbe au singulier apparaissait je ne lisais plus le reste de la phrase (ou inversement).

If you answered "YES" to the previous question, please explain your strategy here.

Translated Replies

- I rely on a feeling and an impression when it is too fast.
- Depending on whether the verb was plural or singular (but not valid for everything) by looking for the subject and the verbs.
- Identify if the noun phrases were plural or singular, and check the correspondence with the verb(s). For this, the determiners and endings of the verbs are sufficient.
- Identifying the singular and plural.
- The types of mistakes were often the same: wrong verb conjugations or inconsistent words in the middle of the sentence. By paying attention to this, I tried to get the answer when the sentence was running too fast.
- Spelling.
- Just looking at the number of plural subjects and plural verbs.
- I would focus on the subject of the sentence, which was quickly falsified or validated by the incoming verb As soon as the error arrived, I stopped reading the sentence. If no error, complete reading.
- By identifying the subject and the agreement of the verb, and sometimes as soon as a nonsense appears in the sentence.
- Read the two nouns and remember if they are plural or singular, then see if the verbs agree.
- I would try to spot the singular and plural and the subjects to see if they matched.
- I was paying attention to the plural/singular agreement by concentrating on the subject and the verb as soon as there were nouns in the plural, as soon as a verb in the singular appeared I didn't read the rest of the sentence (or vice versa).

On a scale of 1 to 9, how natural were the sentences?

3.95 ± 1.78

What was the most important part of the sentence to solve the task?

Options:

- The subject.
- The verb.
- Both

- The subject: 9.5%
- The verb: 28.6%
- Both: 61.9%

On a scale of 1 to 9, how difficult was the experiment?

5.23 ± 1.57

On a scale of 1 to 9, how long was the experiment?

5.47 ± 1.29

On a scale of 1 to 9, how well do you think you did on the task?

5.61 ± 1.61

Were all sentences of the same length?

- 33.3% of the participants replied YES.
- 66.7% of the participants replied NO.

Which was the easiest, and which the hardest presentation?

Easiest SOA:

- 53.63% :375ms
- 26.13% :250ms
- 21.05% :500ms

Hardest SOA:

- 84.2% 125ms
- 15.7% :250ms

Chapter 4

Attractor proximity effects in subject-verb agreement.

Abstract

Formal linguistic theory postulates that language processing is rooted in a uniquely human ability to generate recursively, nested symbolic representations. This ability presupposes a structural encoding of linguistic sequences (structural processing). Nevertheless, alternative explanations have been proposed, in which no structural bias is required (linear processing).

In this study, we revisit the well studied psycholinguistic phenomenon of subject-verb agreement, in the presence of an intervening noun called an attractor. Importantly, we introduce a minimal set of experimental conditions, in which we modulate parametrically the linear distance of the attractor from the verb. Based on evidence from non-linguistic sequence processing and recent studies in human language processing, we sought to identify effects that can be attributed to calculations of transition probabilities between the attractor and the verb. Operations attributed to low-level statistical regularities, such as transition probabilities at the bigram level, could be attributed to linear, rather than structural processing.

We report behavioral results from an online, forced-choice, violation-detection experiment. Additionally, we analyze the behavior of transformer language models, presented with the same set of stimuli. Our results show a clear modulation of the behavioral index by the attractor distance. Notably, the distance effect modulated differently the performance of the human and the artificial system. When the attractor reached the vicinity of the verb, human subjects performed better in detecting violations, whereas the performance of the models reached chance level.

Importantly, our analyses showed that the distance effects can be traced to the inflectional covariation of the head noun and the attractor, and not from the dependency between the attractor and the verb. Furthermore, our results in agreement with the cue-based model of attraction, as this model predicts that the memory representation of the distant attractor would fade away and thus will not lead to similarity based retrieval-interference.

Thus, these results corroborate our previous findings and point to a language processing system driven by structural operations and robust to low-level statistical regularities such as transition probabilities.

4.1 Introduction

On the surface, language appears to be linear, as humans read or hear words one after the other. However, according to formal linguistic theory, there exists an underlying structure that governs language processing [Chomsky, 1957, Rizzi, 2004, Dehaene et al., 2015, Vigliocco and Nicol, 1998, Hauser et al., 2002]. The conception of linguistic structure is a hypothetical notion, nevertheless perceived as a standard in the linguistic community [Uddén et al., 2020]. Even though, there exist both behavioral [Shi et al., 2020, Coopmans et al., 2021] and neural [Nelson et al., 2017, Pallier et al., 2011] accounts for structural operations, alternative interpretations have been proposed in which no structural presupposition is required [Frank and Bod, 2011]. According to the latter view, linguistic phenomena can be explained by statistical relationships at the word level. The discrepancy between the two views has led to a decade-long debate on linear versus structural effects in language processing [Haskell and MacDonald, 2005, Ding et al., 2015, Willer Gold et al., 2017, Arana et al., 2021].

In this study, we tackle this debate by revisiting the subject-verb agreement phenomenon, a phenomenon traditionally attributed to core syntactic operations [Molinaro et al., 2011a]. A popular way to contrast structural operations and linear, word-order effects is the analysis of subject-verb agreement errors in the presence of an intervening noun, called an attractor. Attraction effects in number agreement have been studied extensively in humans [Bock and Miller, 1991, Shen et al., 2013, Tanner et al., 2014, Hammerly et al., 2019, Paape et al., 2021, Sinha et al., 2021] and neural language models (NLMs) [Linzen et al., 2016b, Finlayson et al., 2021, Lakretz et al., 2019b, Jumelet et al., 2019].

In this project, we revisit the subject-verb number agreement phenomenon, and introduce a parametric manipulation of the attractor distance (Figure 4.1). Additionally, we include a baseline condition where the subject-verb distance is the same, but no number-carrying word is introduced within the embedding. We thus provide a minimal triad of experimental conditions (Table 4.1 that aims to disentangle structural from linear mechanisms, by contrasting operations directly ascribed to each mechanism. Figure 4.1 summarizes the main factors of the design.

We assume that the grammatical configuration of the sentence is controlled by structural operations (*Violation* effect). Similar to our previous work, we define two additional effects. The *Congruency* effect stems from the inflectional covariation of the two nouns. Furthermore, based on the codification of sequence processing proposed by Dehaene et al. [Dehaene et al., 2015], we attribute the *Transition* effect to transition-probabilities between the attractor and the noun. This hypothesis stems from the predictive coding framework [Friston et al., 2021] and received recent support by a behavioral study that demonstrated effects of local statistics, at the bigram & trigram level, on the reading time of the participants [Goodkind and Bicknell, 2021]. Lastly, we analyze the behavior of two language models in a comparative way, given the known sensitive of computational models to statistical relationships.

4.2 Methods

Participants Fifty-four native speakers of English took part in an online experiment. The experiment was advertised on social media (Twitter and Facebook) and through several mailing lists, and was targeting native English speakers, although anyone could participate. People interested to participate could simply click on the provided link, read and accept a written consent, in which they declared not to be legally minor. Participants were informed that they could withdraw from the experiment at any moment by simply quitting the webpage, in which case no data was collected. The procedure and the consent were approved by the local ethical committee (Université Paris-Saclay, ref. CER-Paris-Saclay-2019-063).

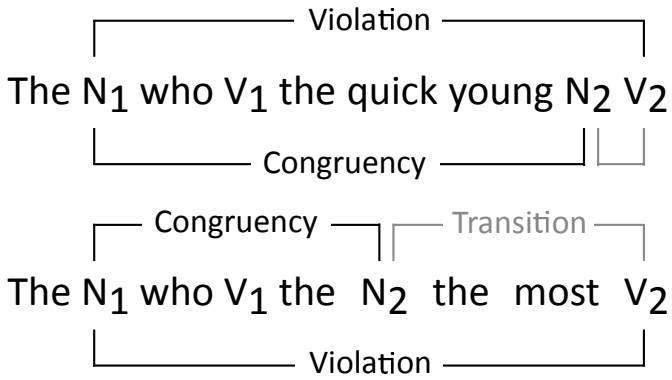


Figure 4.1: **Experimental design:** Our design seeks to contrast two different types of processing in language comprehension. The main effect of *Violation* refers to the dependency that controls the grammatical configuration of the sentence and is used as a proxy into structural processing. The *Congruency* factor refers to a dependency realized between two non-structurally related words (Head:N_1 and attractor:N_2). The *Transition* effect is a dependency realized between the attractor (N_2) and the verb (V_2). As a baseline, we include a condition with the same number of words but with no noun between N_1 and V_2. Table 4.1 summarizes the experimental conditions and the full material is available in supplementary materials.

Condition	Sentence	Violation	Congruency
Proximal Attractor	The doctor who loves the careless young nurse climbs.	✗	✓
	The doctor who loves the careless young nurses climbs.	✗	✗
	The doctor who loves the careless young nurse climb.	✓	✓
	The doctor who loves the careless young nurses climb.	✓	✗
Distal Attractor	The doctor who loves the nurse the most climbs.	✗	✓
	The doctor who loves the nurses the most climbs.	✗	✗
	The doctor who loves the nurse the most climb.	✓	✓
	The doctor who loves the nurses the most climb.	✓	✗
Baseline	The doctor who walks fast but rather clumsily climbs.	✗	-
	The doctor who walks fast but rather clumsily climb.	✓	-
Filler (Number)	The doctor who likes the nurse climbs.	✗	-
	The doctor who like the nurse climbs.	✓	-
Filler (POS)	The doctor whom the nurse likes climbs.	✗	-
	The doctor whom likes the nurse climbs.	✓	-

Table 4.1: **Conditions & stimuli of the design:** The main linguistic construction used in the design is a subject-relative modifier. We present three distinct conditions and two types of filler-trials. The main conditions are separated, first, by the presence or not of an attractor, and subsequently by the attractor-verb distance. The main factors of *Congruency* & *Violation* are presented in Figure 4.1

.

Early piloting suggested that some participants would notice the structure of the trials and purposefully ignore the middle of the sentences. To identify these participants, we made use of the filler trials: any participant whose answer to fillers was not significantly different from chance (binomial test, null hypothesis $p_0 = .5$) was rejected. In this way, we tried to address a possible strategy confound that might be emerging due to the nature of our experiment [Pearlmutter et al., 1999]. Finally, we rejected any participant whose success rate was below 70% on the main task. Overall, we rejected 20 participants, and 34 were analyzed.

Experimental procedure The experimental procedure started with a consent form, then a series of questions on demographic aspects and on subjective self-evaluations, and finally, the instructions. Then, participants were presented with sentences in a rapid-serial-visual-presentation (RSVP) manner with a fixation cross between words, in white on black background, using a presentation time of 200ms and an SOA of 366ms (22 frames on a 60Hz monitor), and could answer at any point by pressing the left or right key (key randomized across subjects, specified in the instructions). Participants received auditory and visual feedback with each trial: green fixation and upward tune (correct), or red fixation and downward tune (incorrect). At the end of the experiment, participants answered a few extra questions about the experiment, then saw their overall score and were invited to share the experiment on social media.

Stimuli First we generated sentences from a fixed lexicon and simple construction rules, yielding a very high number of sentences, part of which were highly improbable. Then we filtered based on a GPT3 language model perplexity. We only kept sentences that had an overall perplexity between the median and median $+2std.$ This left us with a homogeneous pool of sentences from which we sampled stimuli for the experiment. To encourage diversity during sampling, we rejected sentences that shared more than 5 (out of 9) words with an already sampled sentence. We sampled 5 sentences for each condition (baseline, distal congruent, distal incongruent, proximal congruent, proximal incongruent), subject number (singular, plural), and presence of violation (yes, no), as well as 32 filler sentences, totaling 132 sentences. The same set was presented to each subject and neural models. The lexicon was balanced for low-level features such as word length and unigram frequency.

Questionnaire Participants were asked to specify whether they were a native speaker of English, what other language(s) they were native speakers of, their gender, age and highest degree obtained. We also asked them, on a scale from one to 10, how calm their environment was at the time of the experiment.

Instructions Participants were given a description of the task (see appendix for the exact wording) which indicated (i) what a rapid serial visual presentation looks like, (ii) what the task was, and which button they should use to answer, and (iii) three examples of grammatical as well as three examples of ungrammatical sentences. The experiment started with seven training sentences, independent of our stimuli, in order for the participants to get used to the procedure, then participants were instructed that the main body of the experiment started, and they went through the sentences in random order.

Neural Networks We also tested with the same set of sentences two transformer models downloaded from HuggingFace [Wolf et al., 2020]: a causal GPT3 language model [1] [Brown

[1] https://huggingface.co/EleutherAI/gpt-neo-1.3B

et al., 2020] and a Text-To-Text Transformer (T5) [2] [Raffel et al., 2019] fine-tuned on JFLEG, a grammatical error correction benchmark [Napoles et al., 2017]. To evaluate the GPT3 model, we input it with the sentence up to, and excluding the target verb, and compare the probabilities associated with the token that would make the sentence grammatical and the corresponding one that would make the sentence agramatical (e.g., "climb" vs "climbs" for sentences in Table 1). Thus, for this model we only get an overall performance per condition, but cannot split grammatical and agrammatical sentences. To compute the $T5$ model's performance, we input it with the full sentence, and, *i*) for grammatical sentences, we consider a correct answer if the output is identical to the input, and *ii*) for agrammatical sentences, we interpret a correct answer solely if the only change in the output is to fix the agreement error on the final verb.

4.3 Results

We call incongruent, the trials in which the numbers of N_1 and N_2 disagree (Figure 4.1), following standard psycholinguistic terminology.

Figure 4.2 shows the main effect of the number-bearing noun (baseline), and the effect of its distance to the target (distal, proximal), both in our behavioral experiment and the two Neural Language Models (NLMs). In all cases, the baseline features fewer errors and faster reaction times compared to the embedded-noun conditions.

We also find that errors occur more often in agrammatical sentences than in grammatical ones, irrespective of the condition, replicating the phenomenon of *grammatical illusions* [Phillips et al., 2011], indicating that participants more often accept agrammatical sentences than they reject grammatical ones. We also replicate the *grammatical asymmetry* [Wagers et al., 2009] which characterizes subject-verb agreement: incongruent trials led to higher error-rate in the agrammatical sentences.

Already, these results demonstrate a clear modulation of the behavioral index by a non-structurally intervening noun. To investigate the nature of this modulation, we sought to analyze the main factors of our design. The main effect of *Violation* refers to the dependency that controls the grammatical configuration of the sentence (Figure 4.1). Grounded on classical linguistic theory [Rizzi, 2004], we use this effect as a proxy into structural processing. The *Congruency* factor refers to a dependency realized between two non-structurally related words. Due to the factorial nature of our design, the interaction of these two factors, would inevitably lead to the third factor (*Transition*.

Figure 4.3, shows the main factor interaction plots, for both humans and NLMs, and table 4.3 presents the corresponding ANOVA analyses. Importantly, in our analysis, we replicated the markedness phenomenon [Bock and Miller, 1991, Wagers et al., 2009], according to which, attraction effects surface only when the attractor is plural. Thus, results shown in figure 4.3 and table 4.3 correspond to sentences filtered for the attractor number, as the main effect of *Congruency* was not significant in the singular attractor case, irrespective of the distance from the verb[3].

We first sought to investigate the modulation of the behavioral indices by the structural effect. The main effect of *Violation* for the human data was significant across all three conditions and both the error-rate and reaction times (RTs). Nevertheless, for the error-rate, the η^2_G factor (explained model variance) was larger by an order of magnitude in the distal attractor case. This illustrates that participants made more errors in judging the validity of the sentence when the attractor was further away, and especially in detecting agrammatical sentences. Simply put, we

[2] https://huggingface.co/vennify/t5-base-grammar-correction
[3] For a demonstration, see figure 4.4

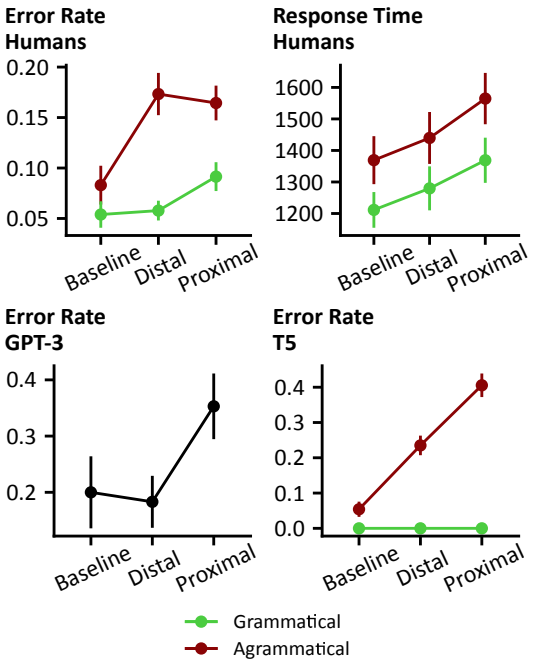


Figure 4.2: Comparing performances of humans (Response Times and Error Rates) and Neural Networks (GPT-3 and T5). Color indicate whether the sentence was grammatical or not. Error bares indicate SEM, over participants for humans and over sentences for neural networks.

Effect	$F_{1,33}$	p-value	η^2_G
Response Time; Distal attractor			
congruency	7.04	.012	.011
violation	11.64	.002	.031
interaction	0.01	.915	$<$ .001
Response Time; Proximal attractor			
congruency	26.90	$<$ .001	.046
violation	11.81	.002	.027
interaction	0.10	.752	$<$ .001
Error Rate; Distal attractor			
congruency	2.29	.140	.013
violation	18.01	$<$ .001	.135
interaction	0.04	.846	$<$ .001
Error Rate; Proximal attractor			
congruency	15.79	$<$ .001	.085
violation	5.33	.027	.040
interaction	0.57	.454	.003

Table 4.2: Four one-way between subjects ANOVAs were conducted to compare the effect of congruency, violation, and their interaction on both Response Time and Error Rates, for both proximal and distal attractors, in sentences were the attractor is plural. Shaded rows with italic text indicate those were the effect was significant at the $p < .05$ level. The lase column provides the η^2_G, an estimator of the variance explained by the ANOVA similar to the r^2 for linear models.

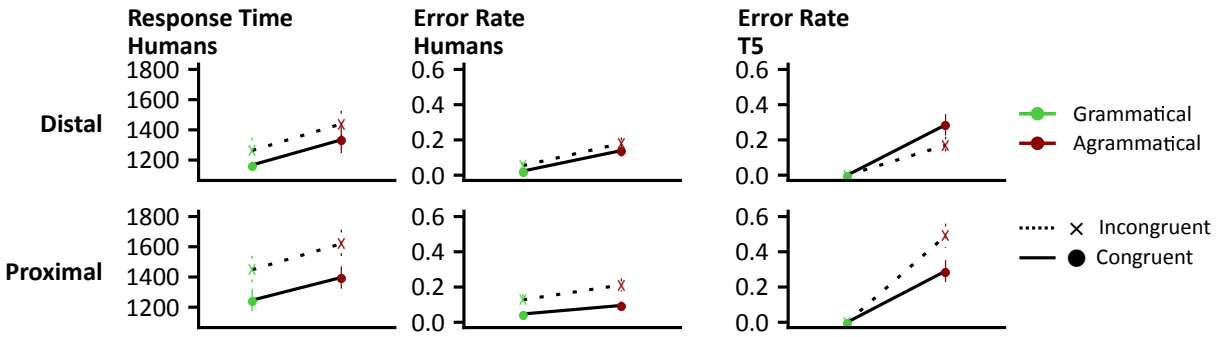


Figure 4.3: Effect of grammaticality, congruency and distance of the attractor on, from left to right: behavioral measures in humans (Response Times, Error Rates), and average error rates in Neural Networks. Error bars indicate SEM, across participants in humans and across sentences in the model.

observed a facilitation effect in the case of the proximal attractor. The distance of the attractor did not affect the magnitude of the effect in the case of the RTs. To elucidate the root of this facilitation, we focused on the *Congruency* effect. The main result of our analysis is, that this effect was only significant when the attractor was close to the verb (proximal condition). This effect illustrates that participants made more errors in the incongruent sentences, compared to the congruent ones. Indeed, figure 4.3 clearly illustrates this. The incongruent trials lead to similar error-rate in both the distal and proximal attractor. Contrariwise, the congruent trials led to a significantly lower error-rate in the proximal attractor condition. Similar to the *Violation* factor, we observed no effect of distance in the RTs, although interestingly, the *Congruency* effect was significant in both conditions.

Importantly, the interaction between the two factors (*Transition* never reached statistical significance, irrespective of the attractor distance.

We analyzed the performance of the models similarly to the human data. Figure 4.2 presents the performance for both models. In particular, the $T5$ model allowed us to split the data based on the grammaticality of the sentences, and thus analyze the data comparably to the humans. Remarkably, the model displayed a super-human performance on evaluating grammatical sentences. In the agrammatical sentences, the mere presence of a number-bearing word led to an increase in the error-rate, similar to the humans. Importantly, we observed a positively correlated modulation of the error-rate as a function of the attractor-distance, unlike with the human data. The $GPT-3$ model did not display the same sensitivity to the bare presence of an intervening element, but crucially, we observed a clear distance effect. Similarly to the $T5$ model, the presence of an attractor in the vicinity of the target verb, led the network to near chance performance.

These results might have a two-folded interpretation. The presence of *grammatical illusions*, in both humans and models, might be informative on the role of training in linguistic performance. In natural language settings, the presence of grammatical sentences surpasses the agrammatical ones. We might therefore be describing a similar training bias between humans and NLMs. This research avenue, although existing, does not fall under the scope of the current project.

The common element in the response profiles of the two models, was the clear sensitivity of the error-rate from the attractor distance. This result illustrates that the sensitivity of the transformer models to statistical regularities is not dependent on the fine-tuning procedure. This fact allows us to use the $T5$ model to draw a comparative picture between the human and the model performance. To investigate the different distance effects observed between the

human and the artificial system, we employed a similar analysis to that of humans, and sought to investigate how the *Congruency* factor modulates the error-rate of the networks. The reduction on the error-rate in the human data was traced to a facilitatory effect of congruency (congruent trials led to fewer errors in the proximal compared to the distal attractor). On the contrary, for the networks, the incongruent trials drove the network to chance level performance, whereas there was no distance effect on the congruent trials. To condense, our results for the proximal attractor case draw the following picture: congruent sentences help participants in performing the task; this is not the case for language models. Additionally, incongruent sentences have a detrimental effect in the models, but no such effect is observed in humans. These observations demonstrate a common sensitivity of attractor-verb distance in both systems, but a fundamental difference as to the outcome of this sensitivity.

4.4 Discussion

In this study, we revisited the classical attraction effect in subject-verb number agreement in humans and neural networks, and sought to dissociate two distinct mechanisms of language comprehension. We introduced a baseline condition where no attraction effects can be realized, and, critically, a distinction between a proximal and distal attractor with respect to the target verb.

We sought to elucidate the effect of attractor proximity to the verb, driven by the hypothesis that low-level predictions at the word level operate in parallel with structural computations in the human system, a hypothesis that has received support by recent neuroimaging and behavioral studies [Shain, 2019, Goodkind and Bicknell, 2021].

Our results draw a clear picture of a shared sensitivity to linear factors between the human and the artificial system, but also a fundamental difference in the effect of this sensitivity. The artificial system operates on the basis of word-level statistics, and is thus driven to chance level performance in the presence of a proximal attractor. On the contrary, the incongruency of the sentence leads to comparable error-rates, irrespective of the attractor distance, for the human subjects. The significant effect of congruency observed in the human data is due to a facilitatory effect of congruency in judging grammaticality, and not an inhibitory effect of incongruency, unlike with the neural networks.

The interplay between the attractor distance and the *Congruency* effect can be decomposed into the following axes.

Incongruent grammatical sentences are harder than congruent.

First, incongruent grammatical sentences become harder compared to their counterpart, congruent ones, only in the case of the proximal attractor. As an example, consider the following two sentences:

a. The doctor who likes the young cute **nurses climbs**.

b. The doctor who likes the **nurses** the most **climbs**.

Our analysis showed that participants made more errors in sentence (a) compared to sentence (b), albeit both being grammatically correct.

Congruent agrammatical sentences are easier than incongruent.

The second observation was that participants parsed the congruent agrammatical sentences easier, with respect to the incongruent ones, only in the case of the proximal attractor. Again,

this can be easier demonstrated with an example. Consider the following two agrammatical, congruent sentences:

c. The doctors who like the young cute **nurses climbs**.

d. The doctors who like the **nurses** the most **climbs**.

Participants made fewer errors in (*c*) compared to (*d*).

Overall, our results point to a modulation of the error-rate by operations occurring at the n-gram level. In the case of grammatical sentences (*a*), we observe the presence of a deviant bigram ('nurses climbs'). We might assume, then, that the *attraction* occurs due to the presence of a locally unacceptable bigram that lures the participants, and causes a disruption in the representation of the subject number. This leads to erroneous answers. Similarly, in agrammatical congruent sentences (*c*), the presence of a locally deviant bigram ('nurses climbs'), facilitates the subjects' performance. It might be that this local deviancy is leading them into judging the sentence as agrammatical, but in this case, the sentence is indeed violated. This might explain the asymmetry observed in our results.

To summarize, the effect of attractor proximity can be decomposed in to two factors. The first is that incongruency leads to more errors in the grammatical sentences, and that congruency leads to fewer errors in the agrammatical sentences. These effects point to a modulation of the error-rate by operations that can be traced to calculations of transition probabilities between the attractor and the verb. Nevertheless, we argue against such operations based on two claims.

First, in our recent work (**??** and in agreement with other behavioral studies [Wagers et al., 2009], we reported significant effects of congruency in non-intervening attractor structures such as Object Relative Clauses (e.g: The doctor that the [nurse like/s] climbs). The n-gram mechanism cannot explain attraction phenomena in this setup, something that the dominant model of attraction (cue-retrieval model, [Wagers et al., 2009]) can do. In this model, a memory mechanism is enabled upon a cue, that retrieves the number of the subject from the memory system and the errors can be attributed to retrieval interference. Under this interpretation, the memory representation of the attractor in the distal condition fades away and therefore does not compete for retrieval. In contrast, when the attractor is in the vicinity of the verb, similarity-based retrieval interference can occur, and thus, attraction effects can only be realized in that condition.

Second, to corroborate this interpretation, it is worth noting that our analyses showed that the attraction effects can be traced to the effect of *Congruency*, that is, effects that stem from the dependency between the head noun and the attractor. Importantly, the effect of *Transition* did not reach statistical significance irrespective of the attractor distance from the verb.

4.5 Conclusion

Taken together, our results suggest that human language processing is reigned by structure-based computations, and is robust to transition effects between non-structurally adjacent words. Additionally, our results illustrate a difference between language processing in humans and neural models.

Supplementary Material

4.6 Instructions

The exact instructions given to the participants are provided below. They consisted of three separate pages, participants could go back and forth between pages freely.

Remember that the key/response binding was randomized across subjects, so page 3 provided below only applied to half of our participants, where the other half had a corresponding, flipped association of key and responses.

Page 1

- This experiment is about sentence processing
- You will read sentences on the screen, with the words presented one after the other, at the center of the screen
- Some of these sentences will contain mistakes
- Your task is to find these mistakes

Page 2

Here are a few examples to show what we mean by correct and incorrect. Remember that the sentence will not be presented as a whole, but rather one word after another.
Incorrect examples:

- The boy drink water while listening to music
- The farmer near the two pilot detests boxing
- The athletes that dislike the happy proud banker sings

Correct examples:

- The boy drinks water while listening to music
- The farmer near the two pilots detests boxing
- The athletes that dislike the happy proud banker sing

Some sentences might be a bit weird, like in example 3, but you should always be able to perform the task if you remain focused.

Page 3

You have to look at the cross at the center of the screen, which is always present when there is no word to read. Make sure the luminosity of your screen is high enough for you to read. Then you will read sentences one word after the other, and you have to do the following:

- As soon as you think a given sentence is INCORRECT, please press the → right arrow key on your keyboard
- When the sentence ends, if you think it is CORRECT, please press the ← left arrow key on your keyboard
- You have to answer every time, even when you're not sure, or you feel you don't know. Only after you answer, the following sentence will start. Answer the best you can!

After each answer you will receive feedback: the central cross will turn green if you answered correctly, and red otherwise. If you can, please turn your computer audio on: that way, you will receive feedback with sounds for each trial.
This is the last instruction page. You can go back to the other pages, but when you move forward the experiment ask you to go fullscreen. Then the experiment will start with 5 training examples so that you understand the task.

4.7 Material

The grammatical sentences we used are all provided below, in their singular variant.

1. The actor that dislikes the lawyer the most prays.
2. The actor who dislikes the chefs the most swims.
3. The athlete that loves the vet the most lies.
4. The athlete who hates the farmers the least sings.
5. The athlete who hates the proud funny woman prays.
6. The author that hates the waiters the least smokes.
7. The baker that hates the lazy gentle man cooks.
8. The baker who dislikes the judge the least cooks.
9. The baker who dislikes the kind helpful plumbers lies.
10. The baker who likes the clever happy plumber swims.
11. The builder that drives happily though rather quickly cheats.
12. The builder who dislikes the proud gentle farmer cheats.
13. The chef that dislikes the proud clumsy tailors sings.
14. The chef who dislikes the authors the least cheats.
15. The doctor that hates the careless young teacher lies.
16. The doctor that runs happily albeit rather carefully cheats.
17. The doctor who hates the actor the least cheats.

18. The farmer who fears the clever lazy tailors cheats.
19. The farmer who fears the doctors the least swims.
20. The farmer who likes the builders the least prays.
21. The lawyer that likes the farmers the most swims.
22. The lawyer that runs carefully yet fairly quickly swims.
23. The man that runs carefully though rather quickly lies.
24. The man who avoids the clumsy helpful chef sings.
25. The man who fears the lazy nice authors lies.
26. The man who laughs carefully yet rather quickly smokes.
27. The man who laughs happily though pretty quickly lies.
28. The man who walks carefully although fairly quickly cheats.
29. The painter that avoids the waiter the most prays.
30. The painter who dislikes the nice careless teacher cheats.
31. The painter who loves the helpful friendly judges prays.
32. The painter who loves the young lazy farmers cheats.
33. The plumber that laughs carefully yet pretty quickly prays.
34. The plumber that rides happily although pretty quickly swims.
35. The plumber who fears the lawyer the most climbs.
36. The plumber who talks happily yet rather quickly swims.
37. The tailor that dislikes the cool lazy baker prays.
38. The tailor that loves the clever clumsy bakers cheats.
39. The tailor who avoids the farmer the least prays.
40. The teacher who dislikes the helpful charming builders cheats.
41. The teacher who fears the lawyers the most cheats.
42. The teacher who likes the bakers the most sings.
43. The vet that likes the proud helpful painters cheats.
44. The waiter who avoids the judge the least cooks.
45. The waiter who dislikes the painter the least prays.
46. The waiter who dislikes the proud nice woman swims.
47. The waiter who hates the cool clumsy man swims.
48. The waiter who likes the actors the most smokes.
49. The woman that fears the baker the most cheats.
50. The woman who avoids the gentle lazy waiters prays.

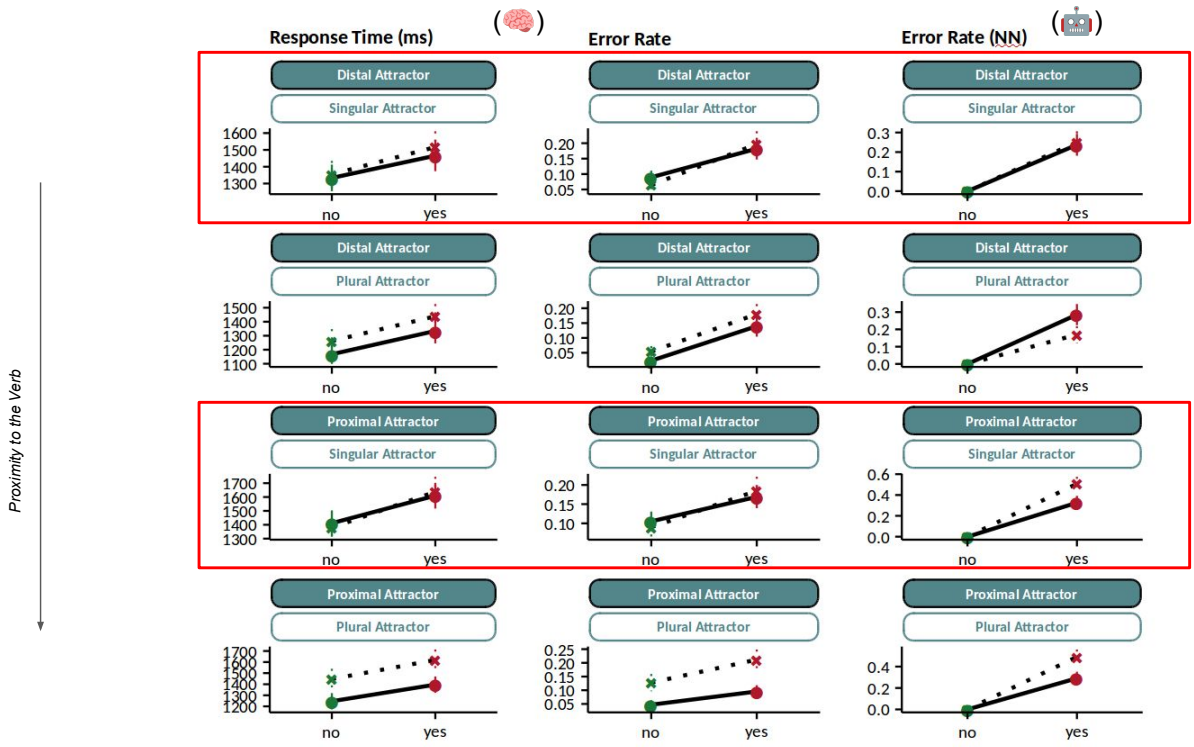


Figure 4.4: The congruency effect only emerges in the case of a plural attractor, irrespective of the distance to the verb.

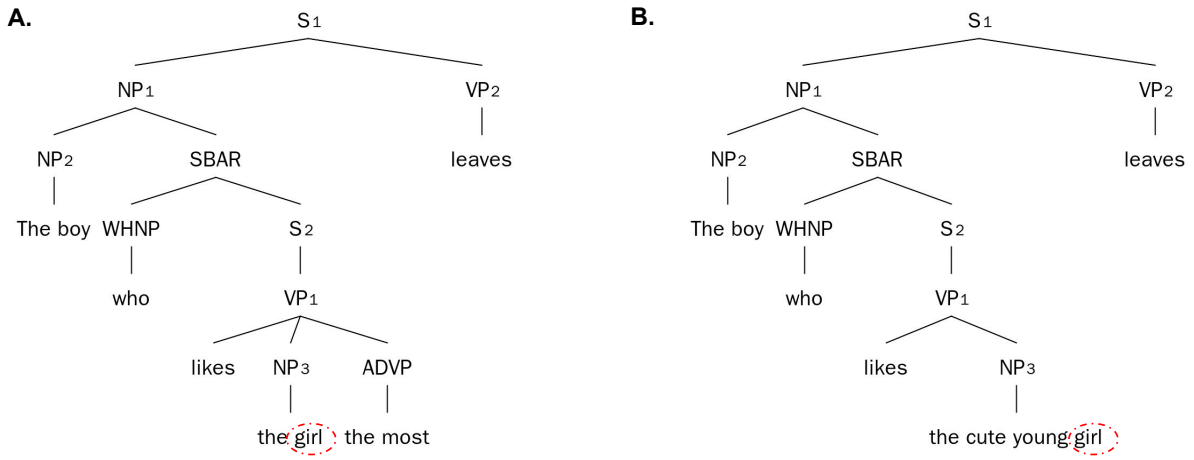


Figure 4.5: The two attractor conditions have similar syntactic representations. In other words, the structural distance between the attractors and the verb is the same in both conditions.

Chapter 5

Discussion

5.1 On the dissociation of two distinct processing mechanisms

Despite years of research, the underpinnings of language processing in the human brain remain a topic that often stirs up heated controversy. There exist two different viewpoints regarding the mechanism that the neural system employs to generate and comprehend language.

One account postulates that language processing is rooted in a uniquely human ability to produce nested symbolic structures (i.e: recursion, [Chomsky, 1957, Hauser et al., 2002, Rizzi, 2004], whereas the second posits that no structural presupposition is required, and that language can be sufficiently explained via probabilistic modelling and statistical learning [Frank and Bod, 2011, Christiansen and Chater, 2015]. We refer to the first mechanism as *structural*, whereas to the second, as *linear*. In this project, we sought to identify correlates of these two mechanisms, using the feature mismatch phenomenon between the subject and the verb. According to formal syntactic theory, this computation is realized by operations stemming exclusively by syntactic or hierarchical relationships. We thus hypothesized that neural or behavioral correlates originating from a factorial manipulation of this dependency can be attributed to the *structural* mechanism. To trace evidence of *linear* processing, we sought for effects stemming from the parametric manipulation of a dependency between the verb, and a noun that does not structurally intervene in the subject-verb agreement configuration (hereafter *attractor*. The backbone of our design was a prepositional phrase modifier with an embedded attractor, such as the following sentence:

The boy near the girls likes climbing.

To pursue this question, we performed a series of two neuroimaging and one behavioral experiment. Across all three experiments, the subjects performed a forced-choice, violation-detection task.

5.2 Brief summary of the results

5.2.1 First study

In this experiment, we utilized two linguistic constructions, a prepositional phrase and an object relative clause. Additionally, we analyzed data from both humans and neural networks. Evidence from non-linguistic sequence processing points to a clear sensitivity of the processing system to simple statistical regularities, such as transition probabilities between successive elements [Bekinschtein et al., 2009b, Dehaene et al., 2015]. In our neuroimaging experiments, we searched for neural and behavioral correlates of a similar, transition-based mechanism between an attractor and a verb that are not structurally related (i.e: “girls likes” in “the boy

who sees the girls likes their looks"). At the behavioral level, we observed a clear influence of the attractor in both constructions, and only for the number feature. Notably, attraction effects were most prominent in the agrammatical sentences, that is, the participants experienced grammatical illusions, but not a-grammatical ones. In other words, we verified, behaviorally, the grammatical asymmetry phenomenon. At the neural level, we failed to decode correlates of a transition-based mechanism, as the classification score remained at chance level. In contrast, the structural effect was traceable for both constructions and features. Contrary to the human recordings, when applying the same analysis in activations extracted from LSTM Neural Networks, the linear effect was fully decodable.

5.2.2 Second study

In our first experiment, we failed to identify neural correlates of a linear processing mechanism, rooted in transition probabilities between two non-structurally adjacent words (attractor and target verb). We speculated that three factors might be responsible for that.

- Transition-based phenomena might be too fast, and therefore dissolved in experimental settings with a relatively slow SOA (in our previous experiment, we used a $500ms$ SOA).
- The morphological complexity of English might not be sufficient to detect these phenomena, given that the inflectional difference between the singular and the plural tense is only denoted by a single letter ("s").
- The participants could have developed task-resolution strategies due to the lack of fillers.

To address the above, we designed a new experiment with multiple SOAs ($125, 250, 375$ & $500ms$), in French, a language with richer morphological variety. Additionally, we included filler trials that contained violations in locations other than the target verb.

In this study, we used the same linguistic constructions as before, but only the feature of grammatical number. Similarly to our first experiment, the decoding of the linear transition-based effect remained at chance level. The structural effect was decodable across both linguistic structures and for the slow SOAs only. Notably, the congruency1 effect reached significance only for the SOA, and only for the prepositional phrase. When taking only correct responses into account, the main effect of congruency was decodable across more SOAs, but only in the object-relative clause.

5.2.3 Third study

Our third experiment, aimed to investigate the effect of linear proximity of the retrieval point, by introducing a parametric modulation of the verb-attractor distance. Additionally, given the known sensitivity of deep learning models to low-level statistical properties, we compared the behavioral performance of humans and two transformer-based models. Attraction effects only emerged when the attractor was in the vicinity of the verb. Furthermore, we verified the two asymmetries that bound the attraction phenomena. Attraction effects occurred mostly in the agrammatical cases (grammatical asymmetry) and only in the case of the plural attractor (mismatch asymmetry). Both the human and the artificial system were sensitive to the attractor-verb distance. Nevertheless, we observed effects that pointed to different directions[1]. Whereas for the humans, we observed a facilitatory interference effect, the performance of the models was severely impacted when the attractor reached the vicinity of the verb.

[1] Given the detection of the grammatical asymmetry, we refer to effects restricted to the agrammatical cases.

5.3 Structural operations dominate the computation of subject-verb agreement

Our results show that during the computation of subject-verb dependency, a non-structurally intervening noun interfered with the agreement computation. These results are in agreement with the existing literature, for both the behavioral [Wagers et al., 2009] and the neural [Tanner and Van Hell, 2014] level.

In our behavioral analysis, and across all three experiments, we observed the phenomenon of grammatical asymmetry. Additionally, in agreement with the work of Wagers, Lau and Phillips [Wagers et al., 2009], we observed significant interference effects in an object relative clause construction, where the attractor does not intervene with the agreement computation. Furthermore, across all experiments, we replicated the phenomenon of grammaticality asymmetry.

Importantly, both at the behavioral, and the neural level we only reported interference effects for the number feature, but not that of animacy. This result is compatible with a syntax-first cognitive model of language processing [Bornkessel and Schlesewsky, 2006, Frazier and Fodor, 1978, Friederici, 2002, Friederici and Weissenborn, 2007]. Based on this approach, during the initial stage of language processing, local phrase structures are built first in an extremely fast manner. Only after this stage has been completed, the parser can continue with building semantic and thematic dependencies. These results strengthen the compatibility of our results with the cue-based retrieval model of attraction [Lewis and Vasishth, 2005, Wagers et al., 2009], as it has been suggested that the error-driven retrieval operations are only sensitive to the number feature [Schlueter et al., 2019]. In particular, behavioral evidence points to a relative 'fragility' of morphosyntactic processing compared to semantic processing of animacy [Stoops and Christianson, 2017].

Additionally, the observed grammatical asymmetry can be primarily explained by the cue-based model of attraction[Wagers et al., 2009, Lewis and Vasishth, 2005, Van Dyke and Johns, 2012, McElree et al., 2003] - although, see [Hammerly et al., 2019]. This model assumes that the agreement is computed based on existing, general purpose memory operations [Jonides et al., 2008], that may retrieve unsuitable elements, given that they share similar features to the ones that should have been retrieved. In the case of the subject-verb agreement, this memory search might retrieve the number of the local number and not that of the controller (subject).

The observed interference patterns indicate, that during the processing of the subject-verb dependency, the number feature of the structurally illicit antecedent perturbed the parser, and affected the resolution of the grammatical configuration. This illustrates that, during sentence processing, structural operations are not completely robust to non-structural interference. Nevertheless, that is not to say that structure does not guide the parser. In fact, our results point to a clear dominance of syntax-based operations. The structural effect of violation was significant, both at the behavioral and the neural level, across all three experiments. Therefore, our results attest to the presence and dominance of structural operations.

The question is now whether, or to what extent, linear operations are also engaged in language processing resolution. Based on both non-linguistic sequence processing [Dehaene et al., 2015] but also studies in computational modelling [Lakretz et al., 2019a], we hypothesized the existence of a transition-probability sensitive mechanism. Indeed, for the neural networks, we were able to decode the transition effect, but in the human data this effect remained at chance level in both experiments.

In our third experiment, we observed a clear impact of the linear distance of the verb from the retrieval point. Congruency effects only emerged when the attractor was linearly

adjacent to the verb. Nevertheless, the main effect of transition[2] did not reach statistical significance. Superficially, the profile of these results seems to point to a sensitivity of the parser to transition probability between consecutive words, a linear parameter. However, an alternative explanation based on the cue-based model exists. When the attractor is further away, the memory representation of the embedded noun declines and thus attraction effects (or effects of retrieval interference) are only detectable in behavior when the verb is presented close to the retrieval point. Given the MEG results of experiments 1 and 2, when no transition effect was found, the latter interpretation is the only one that seems to account for the entire dataset.

Overall, then, across all three studies, the interference patterns can be explained by a faulty-access working memory mechanism. The absence of neural correlates that correspond to a pure transition mechanism is striking given the known sensitivity of the human brain to prediction based on transition probabilities [Friston et al., 2021] and clear evidence of this sensitivity at the non-linguistic sequence level [Dehaene et al., 2015]. It appears, thus, that once a sequence enters the linguistic domain, these mechanisms are of less importance for the processing system, and the structural syntactic representation dominates the picture.

Our results draw an image of a structure-based language processing system, in agreement with formal linguistic theory [Chomsky, 1957, Chomsky, 2009, Rizzi, 2004], where interference from non syntactically licit constituents can be explained via faulty working memory operations at the retrieval stage. Importantly, this interference only emerges for the number feature. Lastly, we demonstrated that language processing between humans and artificial models is different.

5.4 Limitations

In our neuroimaging experiments, we failed to detect neural correlates of a purely transition-based mechanism. Even though we tried to address the possible confounds of our first experiment, it is possible that our studies have limitations that did not allow for the detection of these events.

The subjects performed a forced-choice, violation-detection task. In all of our experiments, we provided immediate feedback at the trial level, with the intention of making the experiment more enjoyable. These types of experiments are repetitive by nature, and the inclusion of the feedback might have transformed them into a sort of linguistic game, thus forcing the subjects into developing strategies towards their resolution. In fact, in our second experiment, we asked explicitly whether the participants *"had found a way of performing the task without reading the whole sentence"*, where 71.4% of the participants replied positively. Importantly, in this experiment, we had included filler trials, hoping to alleviate the development of explicit, task-resolution strategies. Nevertheless, despite the subjects reporting the use of strategies, our behavioral results showed a clear interference pattern. Had the subjects only fully attended to the beginning and the end of the sentences, these interference patterns should not have been detected.

In the same questionnaire, we asked the subjects to rate how natural the sentences were (on a scale from 1 to 10). The mean rating was 3.95 ± 1.78. This might corroborate the notion of a strategy-driven task, and the fact that our analyses might have been reflecting general cognitive functions (such as task resolution), and not natural language processing. However, our results both at the neural and the behavioral level are in line with previous studies where different stimuli and tasks were used [Wagers et al., 2009, Tanner et al., 2017].

[2] Due to the nature of the factorial design, the transition effect can be analyzed as the interaction of the congruency and violation factors

Finally, it is worth noting that the SNR of our analysis might not have been optimal for detecting linear effects, under the assumption that these effects are subtle and difficult to detect.

5.5 Future Research

An approach that could address the above-mentioned concerns, is one that employs a continuous presentation of stimuli in a natural setting, such as for example with an audiobook. Based on a corpus analysis, transitions between grammatically illicit nouns and verbs (such as the attractor-verb dependency), or other locally deviant dependencies can be identified in a natural setting. Thus, the neural correlates of these illicit transitions can be examined in more ecological settings.

The SNR of our M/EEG analysis was limited, therefore, a natural extension of this design, would be one in higher SNR settings, such as intracranial data or even single-cell recordings. Recent work in computational modelling has suggested that the computation of subject-verb agreement is being carried by an extremely sparse mechanism [Lakretz et al., 2019a], thus, single-cell recordings might be the optimal configuration for tracking down such a mechanism. Additionally, to compensate for low-level linguistic factors, future designs should be bimodal (auditory & visual). This allows us to test for generalization across modalities, by, for example, training a classifier on visual data and subsequently evaluating the model performance on auditory. Other than the temporal profile of these computations, of the utmost importance is their localization in the human brain. To that end, the proposed bimodal design should be run in 7T fMRI settings, with an emphasis on single subject analysis.

Additionally, to address whether the results were driven or reflect, task-resolution strategies, it would be interesting to repeat the same factorial experiment, with a much higher proportion of fillers (up to even 50% of the trials) and without an end-of-sentence task.

We sought to identify effects of transitions between non-structurally adjacent elements. For both the animacy and the number features, we failed to trace down such effects. It would be interesting to attest such a claim for the remaining two ϕ-features, person and gender [Molinaro et al., 2011a]. Therefore, a similar design that factorially manipulates these two features, paired with the same type of analysis, would be an exciting future research avenue.

Finally, to assess whether the interference patterns that emerged in our last experiment could be attributed to transition probabilities, it would be interesting to repeat the same exact experiment, but with jabberwocky stimuli. Should the same patterns emerge, we can certify that interference occurs due to faulty memory operations at the morphosyntactic level.

5.6 Concluding summary

In a series of three experiments, we attempted to trace down correlates of two distinct language processing operations. Our results show that once linguistic items enter into the language system they are dominated by structure-sensitive processing, and are largely robust to low-level transition effects. Interference effects were observed only for the feature of grammatical number, and both at the behavioral and the neural level. These interference patterns can be explained by the cue-based retrieval model of attraction.

Appendix A

Appendix for Chapter 2

A.1 The attractor-target transition validity modulates the decodability of violation.

In the following, we present additional analyses on the data presented in chapter2. Figure A.1 shows the modulation of the violation decodability with respect to whether, or not, a valid transition could be realized between the main noun of the sentence and the target verb. This contrast allows us to examine the modulation of a structural effect by surface-lever statistical relationships.

For example, a valid transition coincides with an acceptable bigram in the case of the PP-construction (Panels A & C). Consider the following two sentences, the classification of which corresponds to the discrete line of panel A.

a. The boy near the **girls likes** climbing. *Grammatical-Deviant transition*

b. The boys near the **girls likes** climbing. *Agrammatical-Deviant transition*

On the contrary, the continuous line of panel A corresponds to classification of sentences similar to the following, where a valid transition can be realized between the attractor and the target verb.

c. The boy near the **girl likes** climbing. *Grammatical-Valid transition*

d. The boys near the **girl likes** climbing. *Agrammatical-Valid transition*

Figure A.1 illustrates that in the case of local mismatch between the attractor and the target verb, the linear model reaches higher performance levels. Simply put, it appears that there is an additional processing cost introduced by a feature mismatch between the attractor and the target, at the n-gram level. Importantly, even though the examples corresponding to the PP-Number case draw the picture of a relationship realized at the bi-gram level (see sentence *a*), this is not true for the ObjRC structure. Consider the following sentence pair, corresponding to the discrete line of panel B.

e. The **boys** that the girl **likes** leave. *Grammatical-Invalid transition*

f. The **boys** that the girls **likes** leaves. *Agrammatical-Invalid transition*

In the above examples, the attractor is located away from the target verb, unlike the PP-Number construction. Regardless, we observe a similar decodability pattern. This illustrates, that this effect cannot be attributed to transition effects, but rather stems from memory-interference operations, in agreement with our main results.

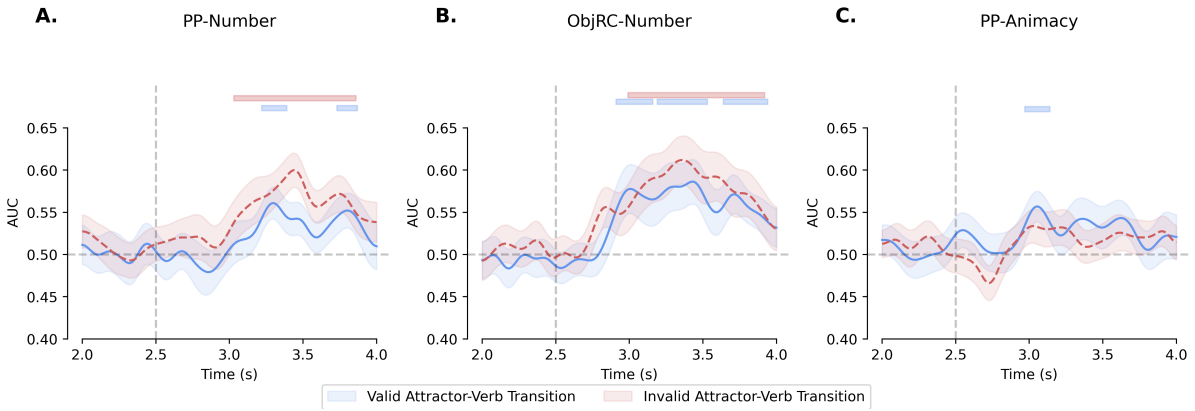


Figure A.1: **Classification of grammaticallity is modulated by the attractor-verb relationship.** Continuous lines correspond to classification of grammaticallity where a valid transition between the attractor and the verb could be realized (sentences a. Vs b. for PP-Number). Discrete lines correspond to sentences where there was a mismatch between the attractor and the target verb (sentences c. Vs d. for PP-Number). The performance of the linear model in the case of a deviant transition is consistently higher and more sustained for the number feature (Panels A & B). This decoding profile is not evident in the animacy case (Panel C). The corresponding lines on the top of each panel correspond to statistically significant time intervals ($p < 0.05$; corrected - spatio-temporal clustering permutation test). Nevertheless, the difference of the two conditions never reached statistical significance in none of the constructions.

A.2 The Grammatical Asymmetry is evident at the behavioral, but not at the neural level.

According to this well-established phenomenon [Wagers et al., 2009], the presence of an attractor leads to more errors in agrammatical sentences compared to grammatical ones. Importantly, this asymmetry can be examined from a different point of view, akin to the classical 'local-global' paradigm for sequence processing.

The investigation of the attraction effect when splitting the design for grammaticality, essentially coincides with the investigation for a pure 'local effect', in terms of the classical 'local-global' paradigm.

For example, the split with respect to attraction and agrammaticallity (figure A.2 - Panel A, right) corresponds to sentences similar to the following:

g. The boy near the **girls like** climbing. *Agrammatical-Valid transition*

h. The boy near the **girl like** climbing. *Agrammatical-Invalid transition*

We first sought to investigate this effect at the behavioral level. Table A.1 summarizes the results of a series of Welch's t-tests, corrected for multiple comparisons, that verify the grammatical asymmetry effect at the behavioral level, and only for the number feature. For the animacy feature (figure A.2 - Panel C), the attraction was examined with respect to the animacy marking of the attractor. The interaction of the number and animacy features was beyond the scope of the current analysis.

Therefore, we verified the grammaticality asymmetry phenomenon at the behavioral level. The presence of an attractor had a significant effect only at the agrammatical cases. We then sought to investigate the neural correlates of this phenomenon. Figure A.3 shows an attempt to decode this asymmetry from the neural recordings. The colors correspond to the grammaticality of the sentences, similarly to that of figure A.3. The red, dashed line corresponds to classification

PP-Number				ObjRC-Number				PP-Animacy			
Grammatical		Violation		Grammatical		Violation		Grammatical		Violation	
p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic
n.s	n.s	0.01	2.57	n.s	n.s	0.002	3.31	n.s	n.s	n.s	n.s

Table A.1: **Behavioral verification of the grammatical asymmetry phenomenon.** Six Welch's t-test with Bonferroni correction were applied to the behavioral data, for each cell of figure A.2.

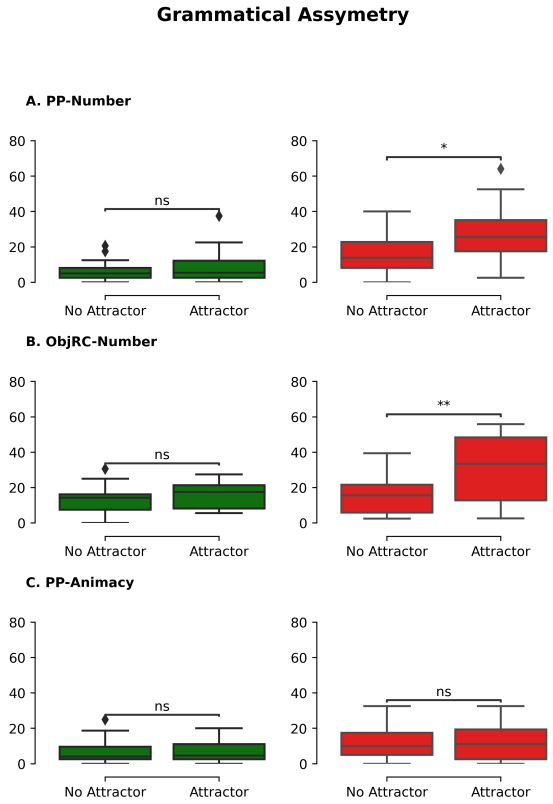


Figure A.2: **The grammatical asymmetry is evident at the behavioral level.** The presence of an attractor had a significant effect only in the agrammatical sentences. We did not observe any attraction effects for the animacy feature.

of agrammatical sentences with and without an attractor (corresponding example: figure A.3 - Panel A. right), whereas the green, continuous line, to the same difference but for grammatical sentences. Should the behavioral data match exactly the neural responses, the decodability of the dashed-red line must have been above chance level for a given time interval, whereas the continuous line should remain at chance level. This inconsistency with the predictions drawn from the behavioral level might illustrate the subtlety of these phenomena, and the need for pursuing high signal-to-noise ratio (SNR) neuroimaging methods for their investigation.

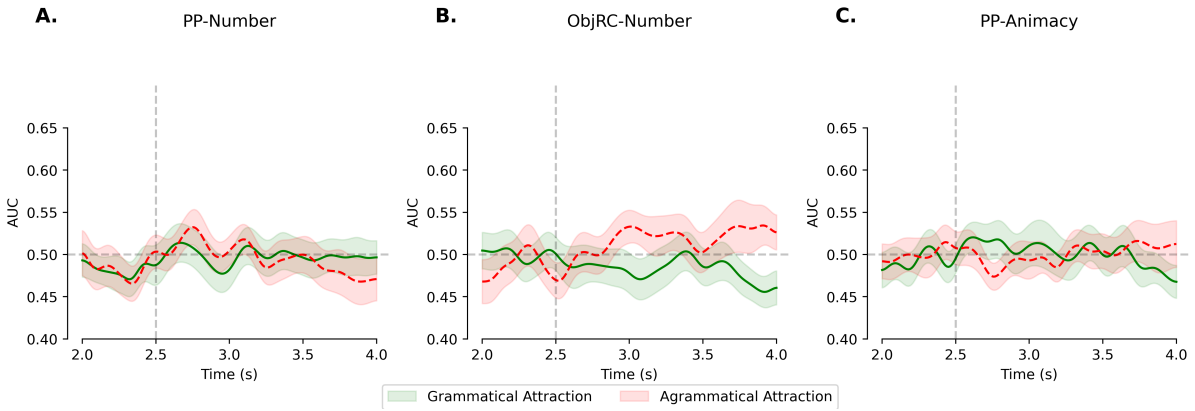


Figure A.3: **No "local effect" at the neural level.** Each panel has a direct correspondence to figure A.2. Even though we observed clear behavioral results, we could not decode them at the neural level.

		PP-Number		ObjRC-Number		PP-Animacy	
		Grammatical	Violation	Grammatical	Violation	Grammatical	Violation
Plural Attractor	Statistic	2.555	3.951	n.s	2.434	n.s	n.s
	p-value	0.01542	0.000326	n.s	0.01972	n.s	n.s
Singular Attractor	Statistic	n.s	n.s	2.112	3.395	n.s	n.s
	p-value	n.s	n.s	0.04068	0.002	n.s	n.s

Table A.2: A Welch's t-test with Bonferroni correction was applied to every cell of figure A.4.

A.3 Neural and behavioral correlates of the markdedness phenomenon.

The markdedness of grammatical number is a well established behavioral phenomenon and states the attraction effect is rooted at the plural noun [Bock and Miller, 1991]. Here, we revisited this phenomenon, by investigating the effect of attraction while splitting the data for both attraction number and grammaticality (Figure: A.4, Table: A.2). The fact that the attractor number leads to different behavioral profiles is often termed "mismatch asymmetry" [Hammerly et al., 2019].

We verified this asymmetry at the behavioral level and only for the number feature, but notably, the error-rate profiles were not identical for the two linguistic structures. For the PP-Number, the results were in agreement with the literature. We observed a very strong, plural attraction effect for the agrammatical sentences ($t(21) = 3.95$, $p < 1e-3$) and a weaker effect on the grammatical sentences ($t(21) = 2.55$, $p < 1e-1$). For the ObjRC-Number construction, the attraction effect was driven by the singular attractor, where the effect was present for both grammatical ($t(21) = 2.12, p < 1e-1$) and agrammatical sentences ($t(21) = 3.39, p < 1e-2$). The plural attractor only had an effect on the agrammatical cases ($t(21) = 2.43, p < 1e-1$). Note that for this construction, the attraction effect stems from an element further away in the linear representation, as demonstrated below.

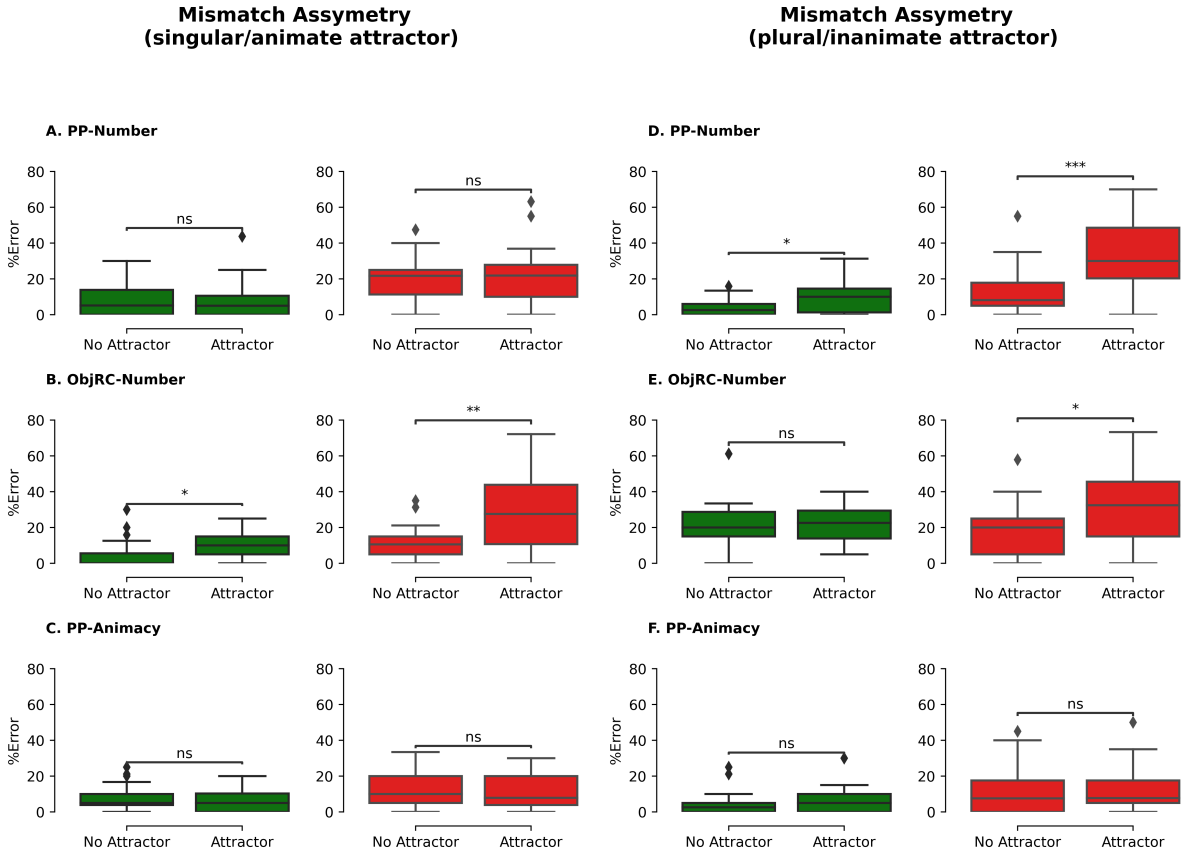


Figure A.4: **The plural attractor drives the attraction effect for the PP-Number construction, but this is not the case for the ObjRC.**

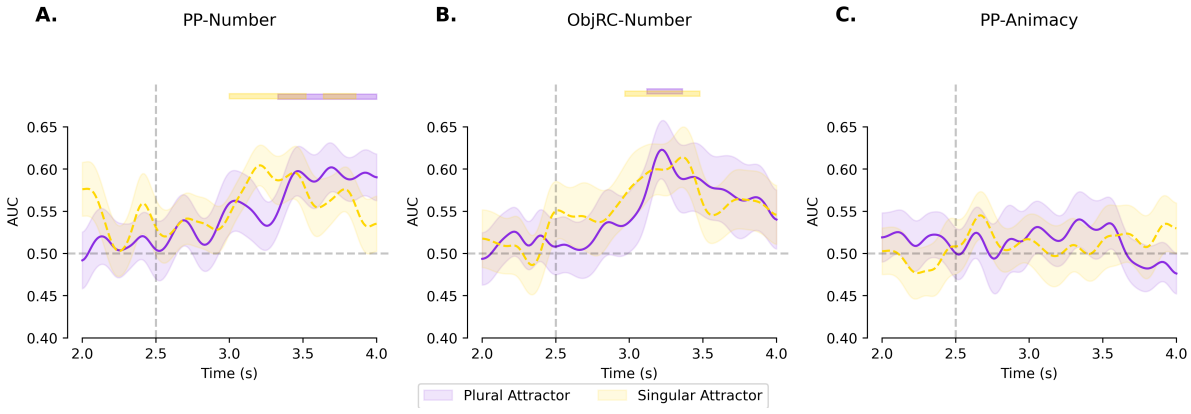


Figure A.5: **The presence of a plural attractor delays the decodability of violation**. This effect is more pronounced in the long-distance dependency where the plural attractor adds an additional $300ms$, compared to a singular attractor.

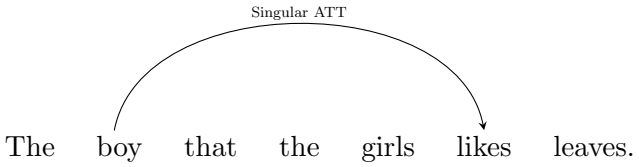


Table A.2 summarizes the mismatch asymmetry results. We did not observe any attraction-like effects when splitting for the animacy of the attractor.

We then sought to investigate this phenomenon at the neural level. Our previous, direct analysis for the corresponding grammatical asymmetry (Figure A.3) did not yield any results, despite the clear behavioral evidence (Figure A.2). We hypothesize that the reduction of the number of trials due to the screening procedure impairs directly the performance of the linear model. For this analysis thus, we employed an indirect approach, in which we further analyzed the dominant effect of violation in the presence of an invalid transition (A.1-dashed line), by splitting the trials for attractor number. Figure A.5 summarizes the results. For the PP-Number construction, the performance of the model was directly comparable for the two attractors, but the temporal profiles of the classification were different. The performance of the linear model reaches significance at exactly $500ms$ after the onset of the target verb in the case of the singular attractor, but requires another $300ms$ in the case of the plural attractor. For the ObjRC-Number construction, we observed as similar pattern, but the difference in the onset of the respective significant decodability was not as pronounced. Nevertheless, the SNR of this analysis appears to be significantly impaired by the further splitting of the trials, which justifies our main analysis approach.

A.4 Generalization of number-violation across constructions.

The design implements the use of two different features (number animacy) and structures (PP ObjRC). So far, we used standard classification techniques to investigate how a linear model can classify between trials of a structure, for a given feature (e.g: Nested-PP-Number). We saw that for the human data, a linear model was able to reach high performance in classifying trials for the main effect of violation for all three constructions.

As a next step, we wanted to examine the ability of the model to generalize across two different axes. First, whether a model trained on data from a given structure can classify unseen data from the same structure, but for a different feature (Generalization Across Feature; e.g: PP-Number $\rightarrow$ PP-Animacy). Next, whether a model trained on data for a given feature can distinguish unseen data for the same feature but from a different structure (Generalization Across Structure; e.g: PP-Number $\rightarrow$ ObjRC-Number). Figure A.6 summarizes the results of these generalizations across the two axes (feature structure) in a 3×3 grid of Generalization Across Time (GAT) matrices ([King and Dehaene, 2014], [Dehaene and King, 2016]) for the violation effect. The diagonals of this grid represent the time-resolved decodabily -measured in AUC-of the linear model, when trained and tested with data from the same construction and feature (e.g: PP-Number). The x-axis of the grid corresponds to the construction on which the model was trained, and subsequently the y-axis represents the testing construction. The dashed lined indicate the significant AUC values. The significance of this generalization was evaluated against chance level using permutation based clustering ([Maris and Oostenveld, 2007]).

The diagonals of this 3×3 grid, correspond to the main effect of violation across the three constructions, locked to the onset of the target word. Our results show that the number feature generalizes well across structure. A decoder trained on the PP-Number structure achieves high performance when tested in classifying neural activity corresponding to the ObjRC structure

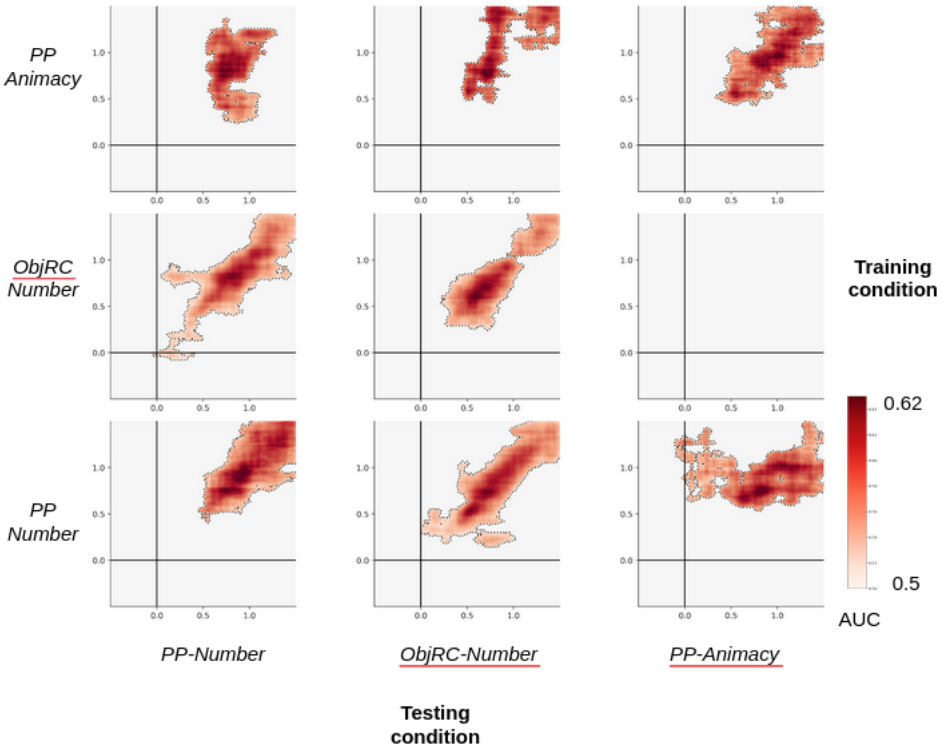


Figure A.6: **Number-violation decoding generalizes across constructions, but weaker generalization from number to animacy violation**. Generalization across conditions and time points for the main effect of violation, for both number (PP- and objRC modifiers) and animacy, measured in Area Under the Curve (AUC). Only significant AUC values are shown ($p < 0.05$; corrected—spatio-temporal clustering permutation test). Dashed contours indicate cluster-level significance. Continuous horizontal and vertical lines indicate target-verb onset. The rows indicate the modifier on which the classifier was tested, whereas the columns the one on which it was trained (i.e: second row, first column: Trained on PP-Number and tested on ObjRC-Number).

and vice versa. The opposite is not true. For example, a decoder trained on ObjRC-Number, fails to reach significance on data from PP-Animacy. These results indicate that the two different features are encoded and processed differently by the neural system.

A.5 Different lateralization for each feature.

The fact that the two features lead to different generalizations can also be seen at their respective topographic plots (A.7). The violation of the number feature leads to left lateralized topographies, whereas the violation effect for the animacy feature led to topographies that highlight a right-frontal representation.

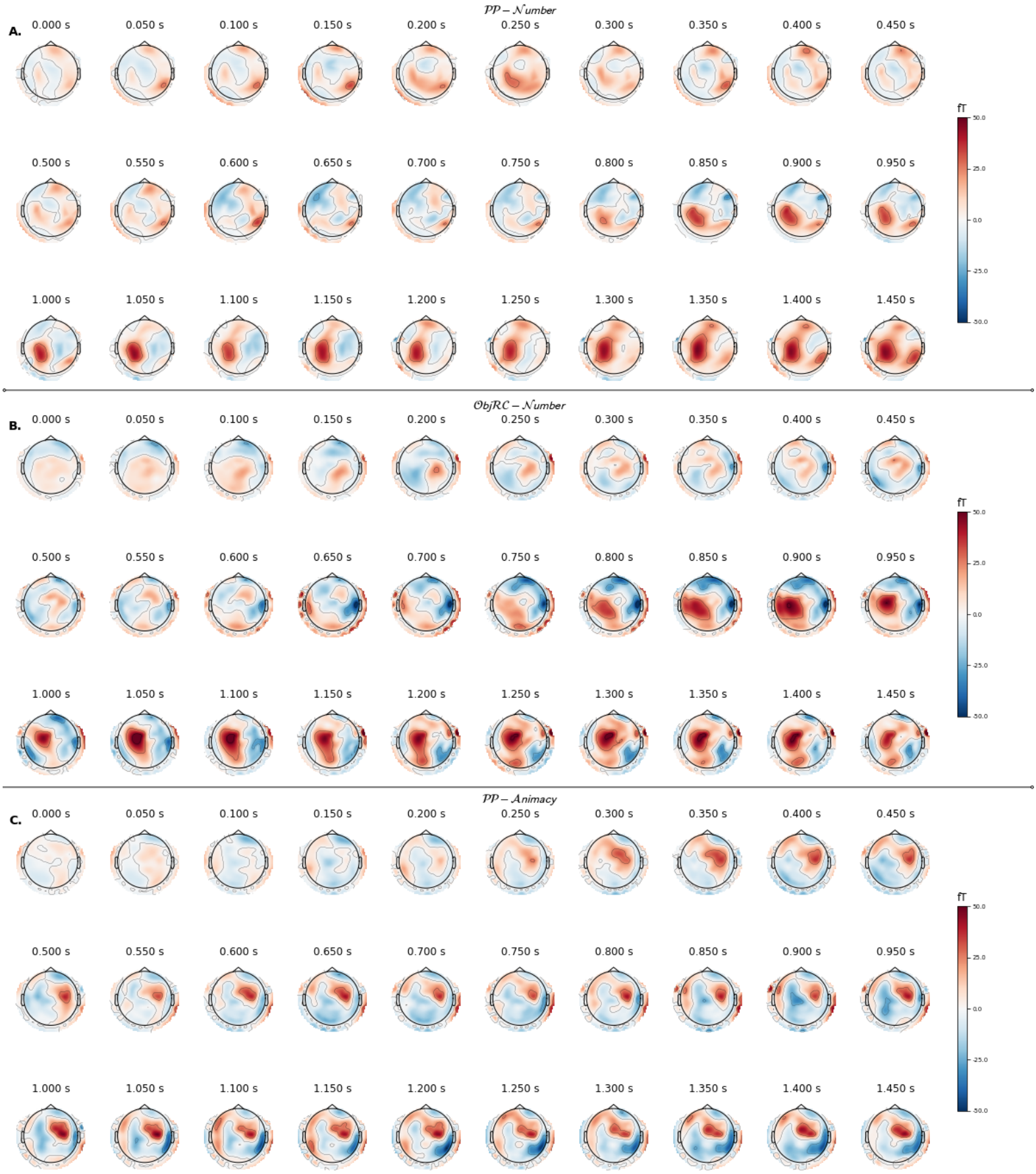


Figure A.7: **Grand average magnetometer topographic plots for the main effect of violation across the three constructions.** Time zero corresponds to the onset of the target word.

Appendix B

Technical work

The following pages showcase a demonstration of a preprocessing pipeline that I co-developed during the first year of my PhD, to identify channels that contain epileptic activity. The pipeline:

- Handles data from different hospitals and recording setups.
- Applies basic signal processing steps (line-noise removal, downsampling, and linear detrending).
- Identifies channels and trials that contain epileptic activity.
- Offers feedback to the user at each intermediate analysis level.
- Delivers flexible solutions to run into multiple workstations irrespective of hardware capabilities of the user and various slower options with lower consumption offered as alternatives.
- Ensures compatibility with Windows and Linux operating systems.

iEEG Cleaning Pipeline

@UNICOG, NeuroSpin

The 4 main Steps for cleaning the iEEG data

1. Detection of hardware artifacts (saturated noise, unplugged electrodes etc) via median thresholding.

2. Detection of Spiking channels.

3. Detection of deviant channels from the Power-Spectrum.

4. Detection of pathological channels based on the detection of HFOs.

General pre-processing steps

1. Downsampling of the data (most iEEG recording systems record at an extremely high sampling rate ~2KHz. It is advised to downsample your data at 1KHz to ease processing and memory allocation).

2. Removal of line-noise and harmonics using a notch filter (unless you're interested in fast-Ripple detection research (>250Hz), a removal of the two first harmonics will be sufficient).

3. Linear-detrending of the data.

4. Prior to analyzing, cut the continuous data into pieces to ease resources allocation.

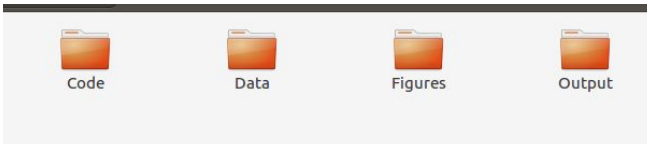
```
% Pre-processing : This is the pre-processing pipeline for the intracranial
% data. The goal of this scrip is to be as generic as possible, integrating
% data from various research centers such as the Houston medical center,
% the Marseille research center etc.
% Written by : Christos-Nikolaos Zacharopoulos @UNICOG 2018
%           christonik@gmail.com
% based on a similar pipeline used at the Stanford University.
% This is a plug-and-play function, to change the research center under
% consideration, provide it as a pair-input in the command window.
% e.g : runPreprocessing('Hospital',{'Houston'})


function runPreprocessing(varargin)
% ---------- BRANCH 1 - SET THE PATHS ---------- %
% In this branch we define the paths for the user.
clc;  close all;
addpath(genpath('functions'));
% Get the data path based on the hostname of the computer in use. The
% variables are unaffected from the OS.
[~,hard_drive_path, elecs_path] = getCore;
% Check whether the parallel computing toolbox is installed - if yes get
% the default number of available cores.
if license('test', 'Distrib_Computing_Toolbox')
   % Get the number of default workers
   numCores = feature('numcores');
   % Open local cluster
   parpool(numCores);
end
% ---------- BRANCH 2 - SPECIFY THE RESEARCH CENTER ---------- %
% In this branch we manunally add the list of patients that corresponds to each
% individual project.
P = parsePairs(varargin);
checkField(P,'processing','quick');
% The other available option is 'slow'/'quick'.
% quick : minimaze file I/O  and visualization to save computing time.
checkField(P,'Hospital',{'Houston'});
```

This is a function with default settings. Every input in this function comes in pairs (this is where the function parsePairs is used.) This is done to increase human readability when it comes to the inputs. For example, to change the processing option, type :

runPreprocessing('processing','slow')

```
inPreprocessing.m | getCore.m | +
    [~, name] = system('hostname');
    % Get the Hostname (Computer ID)
    name = strip(name);
    % Christos - Windows PC
    if strcmp(name,'DESKTOP-4ALPTJB')
      hard_drive_path = fullfile('F:', filesep);
      script_path = fullfile('C:', 'Projects');
      elecs_path = fullfile(hard_drive_path, 'NeuroSyntax2','Data','Houston');
      format compact
      format shortG


      % Fanis - Workstation


    end

  elseif isunix == 1
    % Here, distinguish between multiple UNIX machines
    [~, name] = system('hostname');
    % Get the Hostname (Computer ID)
    name = strip(name);
    if strcmp(name, 'is154105')
      % Christos - CEA laptop
      script_path = fullfile(filesep,'home', 'czacharo','Projects');
      hard_drive_path = fullfile(filesep,'media', 'czacharo','Transcend');
      elecs_path = fullfile(hard_drive_path, 'NeuroSyntax2','Data','Houston');
      format compact
      format shortG
    elseif strcmp(name, 'is150940')
      % Christos workstation - NeuroSpin
      script_path = fullfile(filesep,'home', 'czacharo','Projects');
      hard_drive_path = fullfile(filesep, 'neurospin', 'unicog', 'protocols', 'intracranial');
      elecs_path = fullfile(hard_drive_path, 'NeuroSyntax2','Data','Houston');
      format compact
      format shortG
      % Fosca - Workstation



      % Fernanda - linux Laptop
    end
  end

  % Add the path to the load_settings_params function
  addpath(fullfile(script_path,'Core'));
```

The “getCore” function is used to set the paths. This function is build to set the paths irrelevant of the OS used by the user. Moreover, the user can also specify the hostname of the PC in use, and the paths will be automatically set as long as the file-tree where the code is stored has the following structure:



Once the user specifies the hostname and set the “core” path that leads to the above tree (this can be a hard-drive or the server) the paths are set automatically.

That makes it easy to git-push and work between different people, OSs and PCs under the same OS.

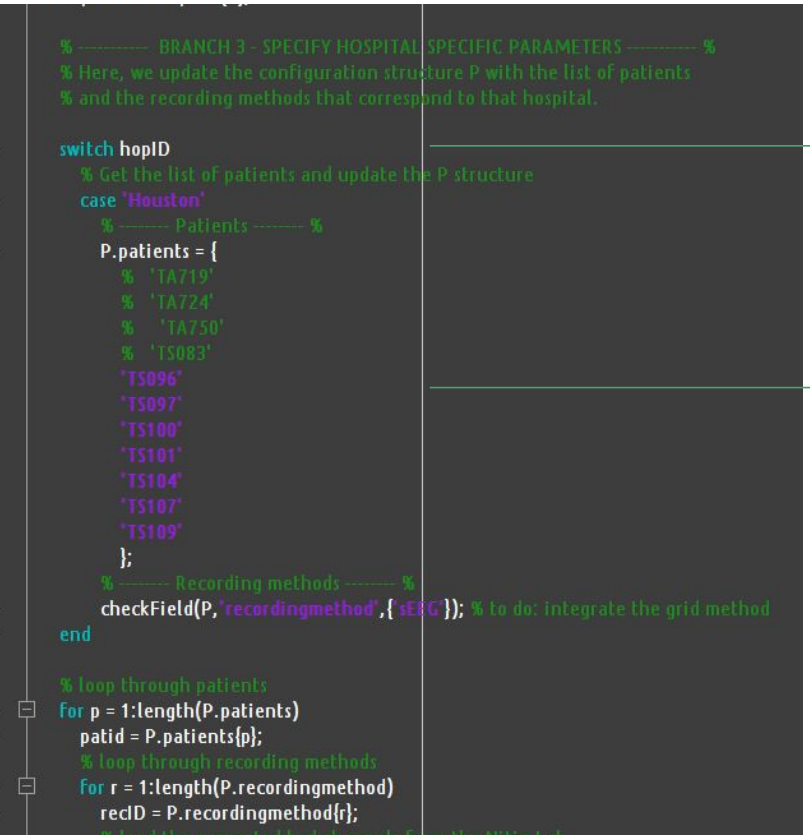
```
  % Check whether the parallel computing toolbox is installed - if yes get
  % the default number of available cores.
  if license('test', 'Distrib_Computing_Toolbox')
    % Get the number of default workers
    numCores = feature('numcores');
    % Open local cluster
    parpool(numCores);
  end
```

Automatically detect whether the parallel computing toolbox is installed. If so, soft-code the number of workers and open the local cluster.

```
  % ---------- BRANCH 2 - SPECIFY THE RESEARCH CENTER ---------- %
  % In this branch we manunally add the list of patients that corresponds to each
  % individual project.
  P = parsePairs(varargin);
  checkField(P,'processing','quick');
  % The other available option is 'slow'/'quick'.
  % quick : minimaze file I/O  and visualization to save computing time.
  checkField(P,'Hospital',{'Houston'});
```

From this point onward, we start to build the configuration structure. This will include the high-level parameters such as the Hospital where we analyzing the data from, the type of processing, etc.

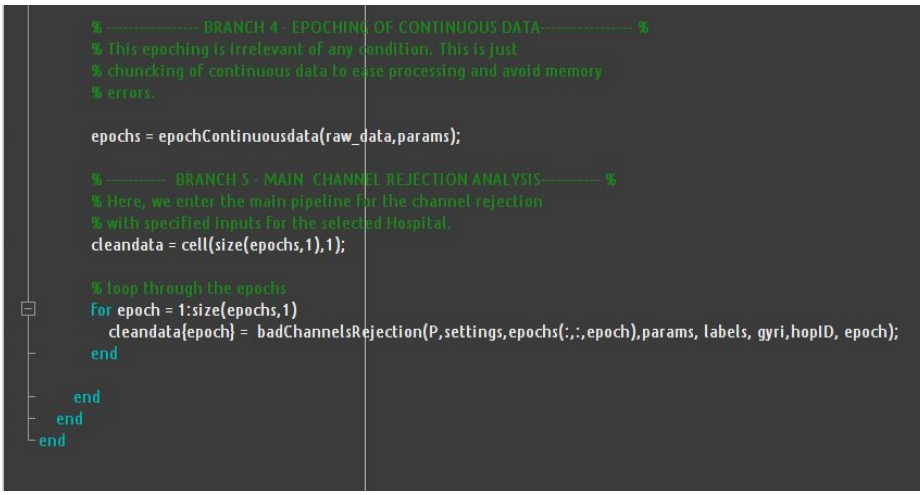
The option (‘processing’,’slow’) allows for visual inspection of the rejected channels at each step. Also, it saves the data from every step on the data path. It is recommended especially for the first time that you run the analysis on a new patient. The option ‘quick’ only informs the user on the number and location of the rejected channels, without showing them.

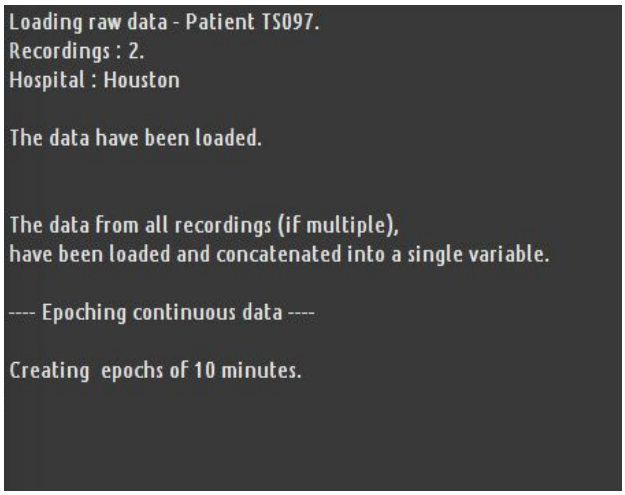


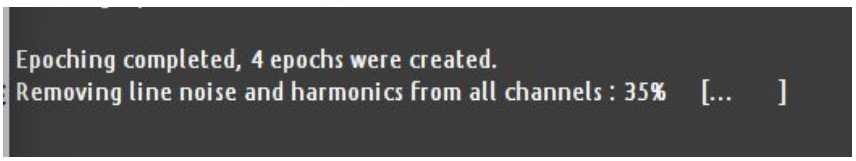
"hopID" stands for "Hospital ID"

Here, we specify the list of patients per hospital. We can either select a single patient or loop through all patients.

The processing for the Grid electrodes has not yet been implemented.







At each step of the process, the user get feedback on the command window.

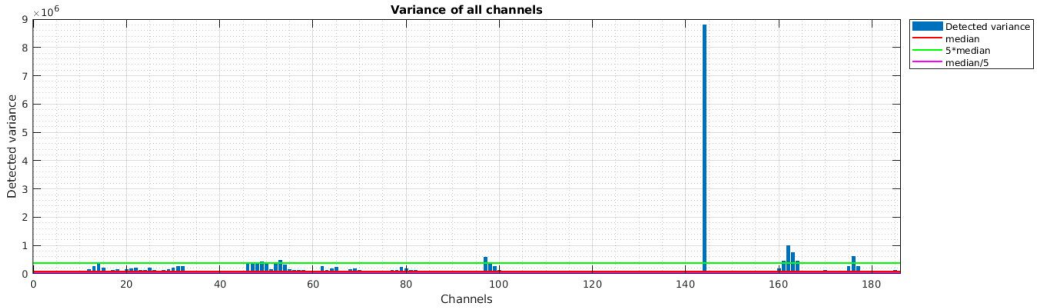
```
switch hopID
  case 'Houston'
    % ---------- STEP 0 ---------- %
    % Non-pathological cleaning steps

    % Create a channel log-file. This will be a logical array where 1
    % will denote a good channel and 0 will denote a bad channel.

    % Initialize variable to hold the filtered data
    channels = size(raw_data,1);
    % Get the duration of the recording
    duration = size(raw_data,2);

    filtered_data = zeros(channels-1,duration);
    % Initialize a logical array where we assume all channels to be
    % good
    allchannels = true(channels-1,1);


    timecount = linspace(1,100,size(filtered_data,1));
    close all;
    textprogressbar('Removing line noise and harmonics from all channels : ');
    % Filter the line noise and the harmonics
    for channel = 1:(channels-1)
      % Do not include the trigger channel
      textprogressbar(timecount(channel));
      % Downsample the data to 1kHz
      wave = downsample(raw_data(channel,:),1,0);
      [wave]= notch(wave, params.srate, 59, 61,1);
      [wave]= notch(wave, params.srate, 118,122,1); % Second harmonic of 60
      [wave]= notch(wave, params.srate, 178,182,1); % Third harmonic of 60

      filtered_data(channel,:) = wave;

    end


    % Remove the trigger channel labels
    labels(end) = [];
    gyri(end) = [];
```

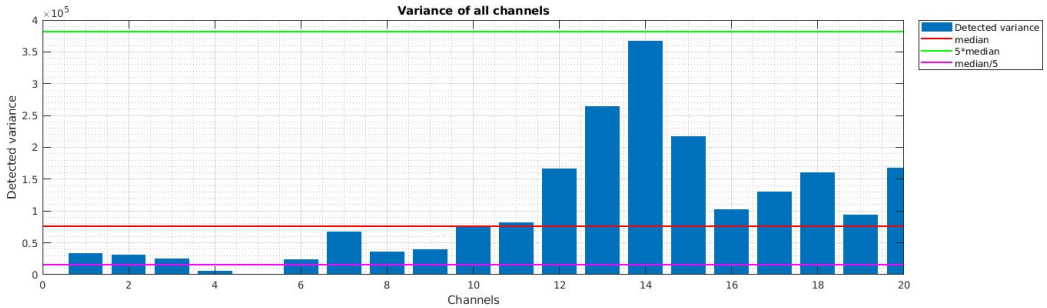
We initialize a channel logical array (in Houston, the trigger is on the last channel, so we exclude it here.)

1 indicates a good channel
0 indicates a bad channel

Filter with the notch filter and downsample here. At this point you can also change the bandwidth of the filter. In the future this will be automated based on where the data come from (Europe or otherwise)

Step 1 : Median thresholding (Rejection based on raw power)





We get the variance of all channels. This returns al. row matrix of dimensions (1 x channels) – we have a single variance value per channel (σ^2/channel).

We then threshold and exclude those channels that exceed an upper and a lower threshold.

Upper threshold = 5*median(var(all_channels))

Lower threshold = median(var(all_channels))/5

```
In total 16 have been removed based on the variance of the all channels.
 The channels have the following labels :

    'LIN4'
    'LIN5'
    'AER3'
    'TP2'
    'TP3'
    'TP4'
    'TP5'
    'TP6'
    'TP8'
    'TP9'
    'TOP1'
    'AH4'
    'AH5'
    'AH6'
    'AH7'
    'PH5'

and are located in the following regions :
    'S_occipito-temporal_lateral'
    'S_occipito-temporal_lateral'
    'G_occipit-temp_med-Parahippocampa…'
    'Pole_temporal'
    'Pole_temporal'
    'Pole_temporal'
    'G_temp_sup-Planum_polare'
    'G_temp_sup-Planum_polare'
    'G_temp_sup-Planum_polare'
    'G_temp_sup-Lateral_aspect'
    'G_temp_sup-Lateral_aspect'
    'Medial_wall'
    'Medial_wall'
    'Medial_wall'
    'Medial_wall'
    'Medial_wall'

Detecting spiking channels.
100%    [..........]
Detection completed.
In total 6 have been removed due to spiking activity.
 The channels have the following labels :
    'AER1'
    'AER2'
    'MBT6'
    'AH3'
    'PH4'
    'PH6'
and are located in the following regions :
    'G_occipit-temp_med-Parahippocampa…'
    'G_occipit-temp_med-Parahippocampa…'
    'G_temporal_middle'
    'G_occipit-temp_med-Parahippocampa…'
    'Medial_wall'
    'Medial_wall'
So far,  22 channels have been rejected out of the total 186.

16 of them have been rejected based on raw power
6 due to detected spiking activity.
```

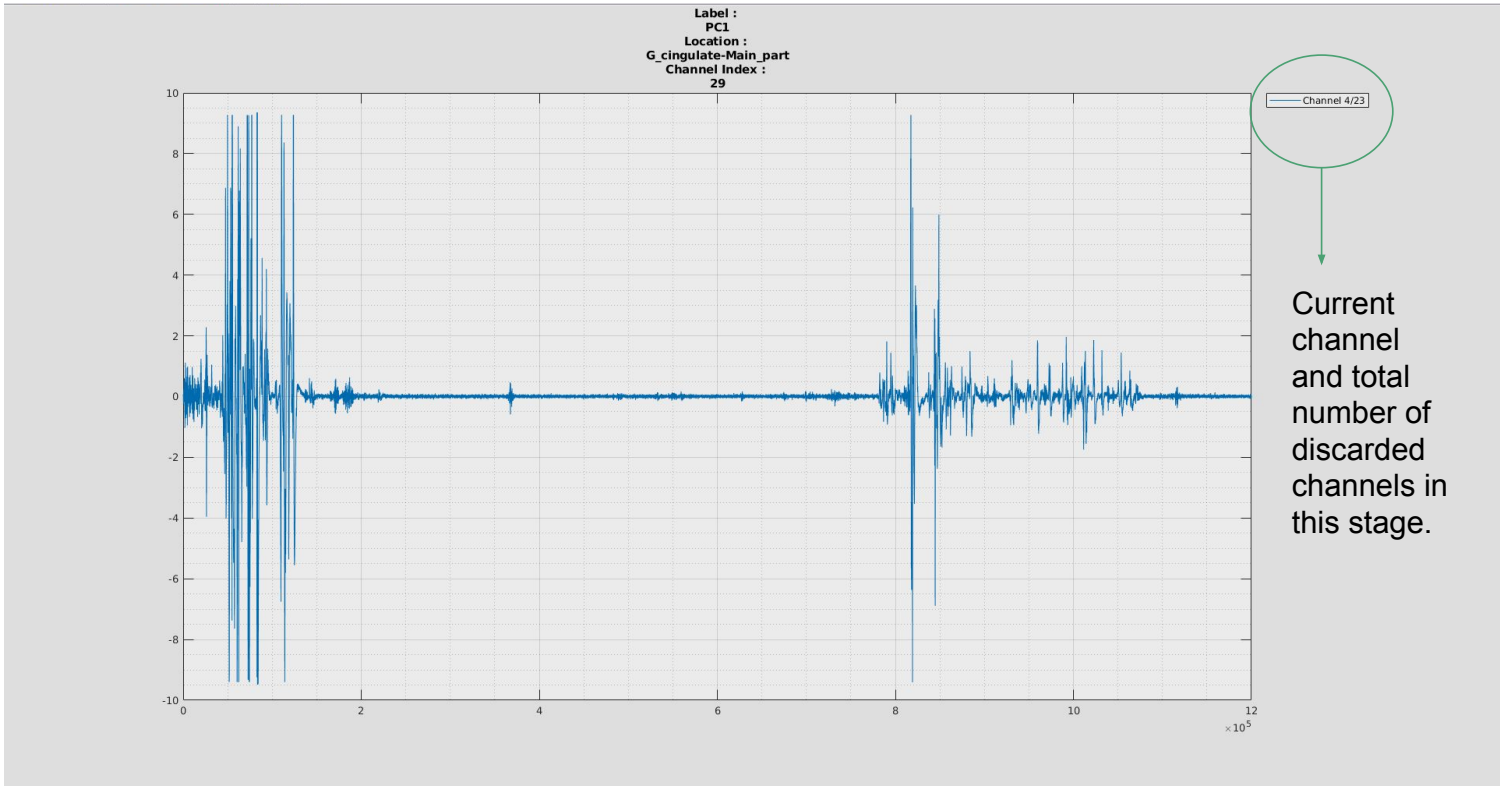
The user receives feedback at each stage of the processing. Here, we see that 16 channels were rejected on the first stage.

The labels of the discarded channels

The locations of the discarded channels

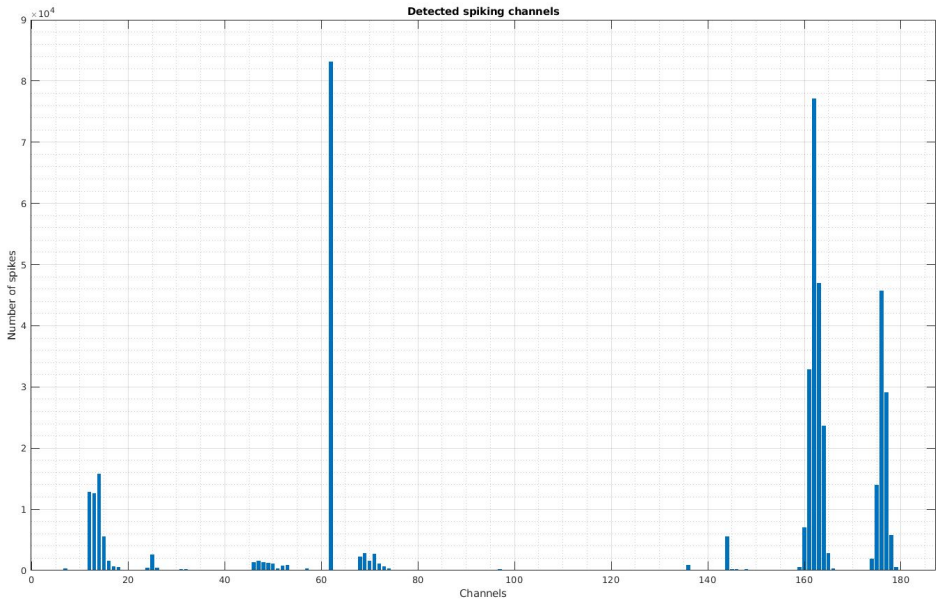
The same holds for each step. In step n.2, 6 channels were rejected

Example of a discarded channel from stage 1.



Current channel and total number of discarded channels in this stage.

Step 2 : Detection of Spiking channels

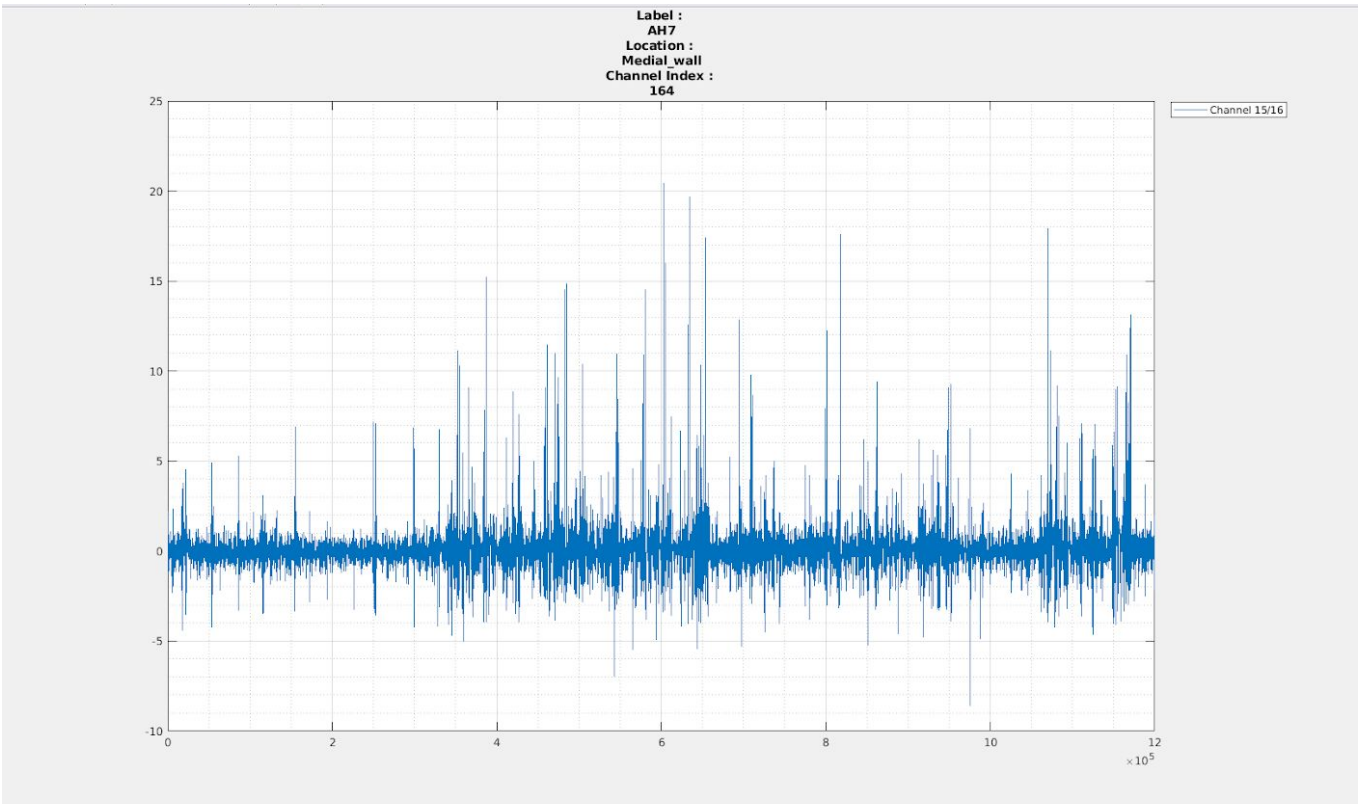


The goal here is to detect rapid changes in the signal (“jumps”).

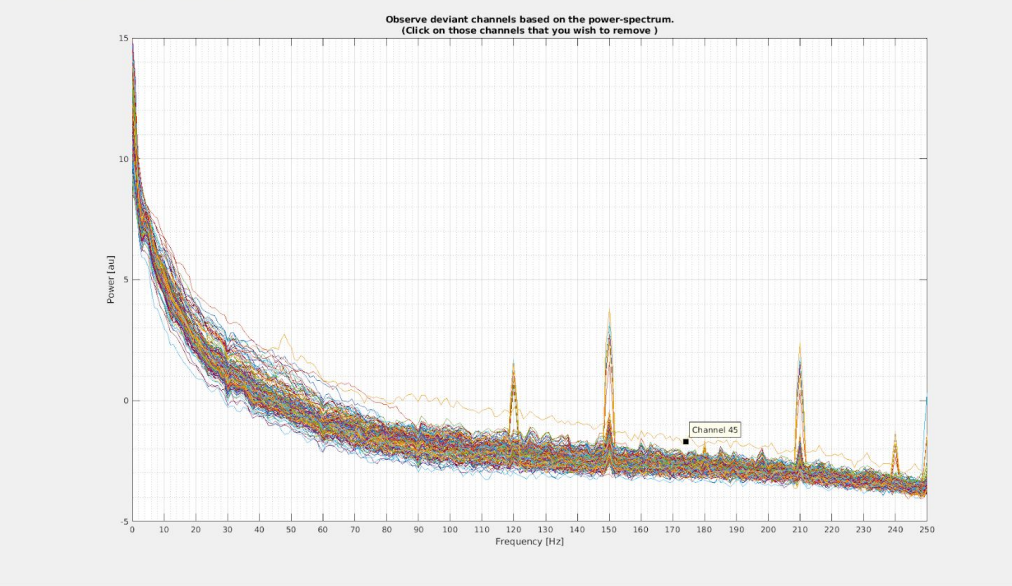
Initially, we set a threshold in mV (80 mV).

We loop through each individual channel and get the difference of two successive points. If this difference exceeds the provided threshold, we call that a spike and
we register the spiking event on the channel.

Example of a discarded channel from stage 2.



Step 3 : Detection of deviant channels from the Power - Spectrum

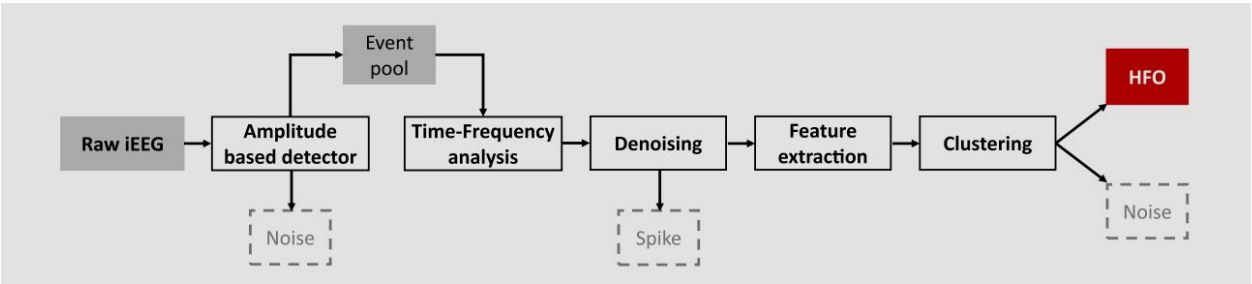


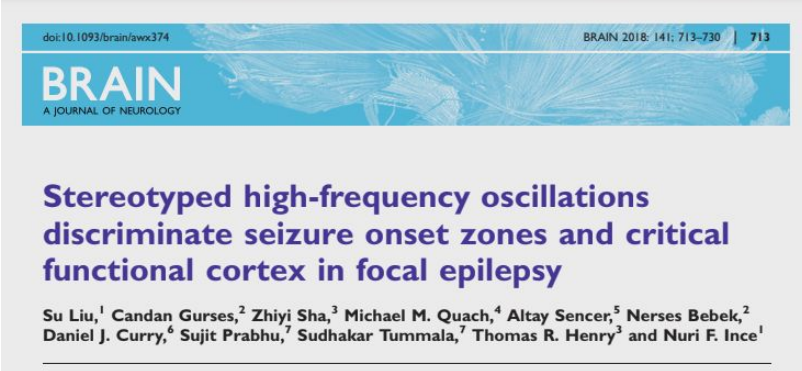
A callback function allows the user to see the index of the deviant channel (in that case, 45).

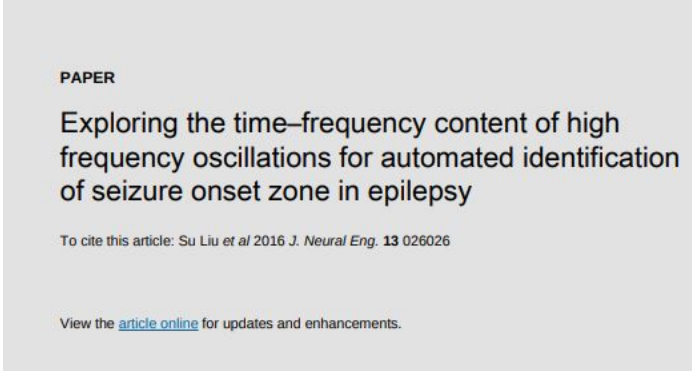
The user can then update the rejected channels on the command window

```
Calculating the Welch's Power Spectral Density
100%    [..........]
Estimation completed
Do you want to add channels for rejection? Y/N :
```

Step 4 : Rejection of channels based on the presence of HFOs







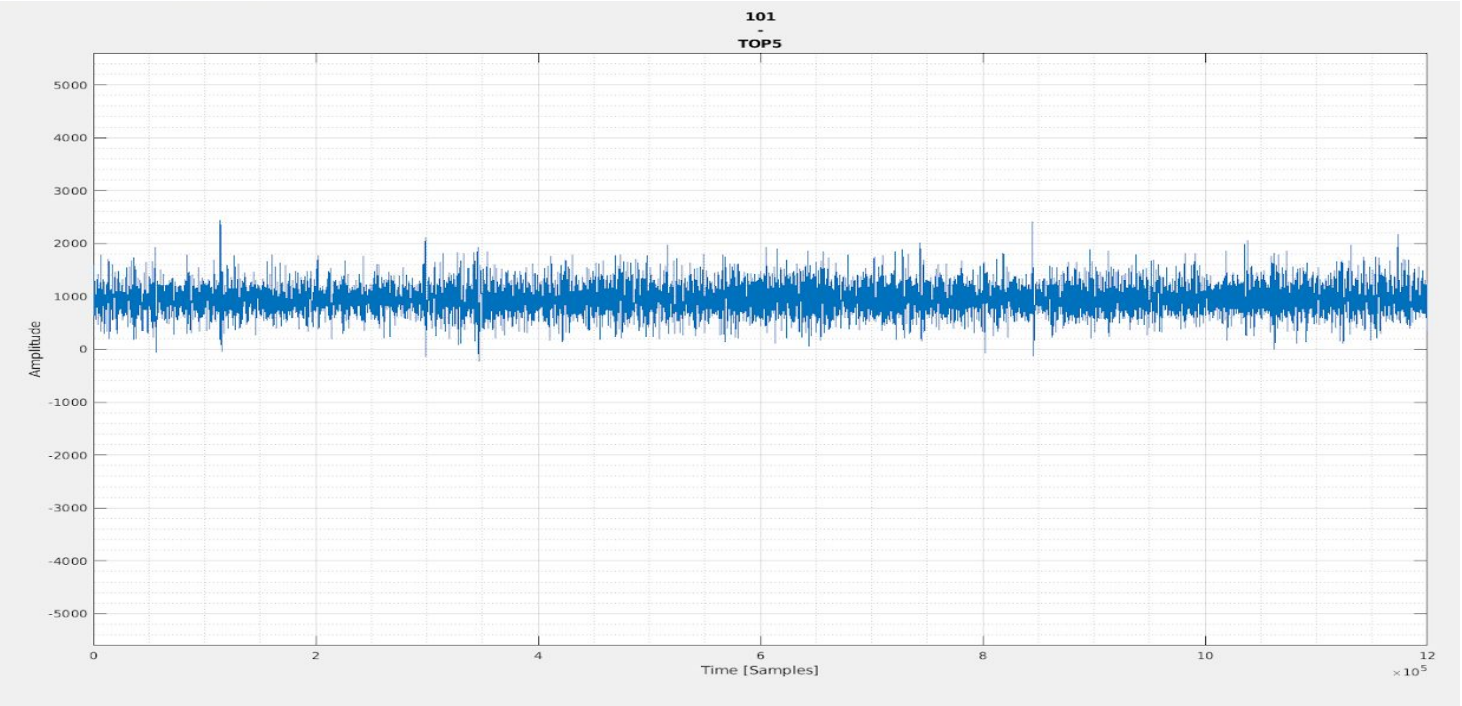
Concluding notes

- The code is not fully vectorized, yet for the largest part, it is.
- To make this code as generic as possible, I've identified several hot-spots where a memory error can occur. At those points, I tried to provide alternative solutions (slower). This code has been tested in Windows (i7,16gb ddr3), Linux (i7,16gb ddr3), Linux (i7,32gb ddr3) and two matlab versions (2017b, 2018a)

```
    catch ('Out of memory. Type HELP MEMORY for your options');
      disp(['The available RAM limit has been reached. ' newline ...
        'Trying to calculate the variance for each channel manually.'])
      % Pre-allocate the variance variable
      dataVariance = zeros(1,size(filtered_data,1));
      for channel = 1: size(filtered_data,1)
        dataVariance(1,channel) = var(filtered_data(channel,:)');
      end
      disp(['The calculation has been completed'])
      switch P.processing
        case 'slow'
          figureDim = [0 0 1 1];
```

- At the end of the pipeline, the user can eyeball the discarded channels, where different colors indicated rejection in different stages.
    - Stage 1 : blue
    - Stage 2 : black
    - Stage 3 : magenda
    - Stage 4 : red
- After applying CAR (Common Average Re-referencing), the user can eyeball the non-rejected channels
- Up to this point, the user has to mentally-note the channels that he/she wants to keep or reject after the visual inspection and manually add them at the editor. I will soon implement an interactive way of doing that from the command window.
- Experiment with the thresholds and find a "Hospital-specific" list of thresholds for each stage.
- So far, this procedure has been implemented at the channel level. I will also implement it at the epoch level.

Example of a non-discarded channel.



Bibliography

[Arana et al., 2021] Arana, S. L., Schoffelen, J.-M., Mitchell, T., and Hagoort, P. (2021). Mvpa does not reveal neural representations of hierarchical linguistic structure in meg. *bioRxiv*. 1, 70

[Armeni et al., 2017] Armeni, K., Willems, R. M., and Frank, S. L. (2017). Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience & Biobehavioral Reviews*, 83(Supplement C):579–588. 5, 13, 52, 53

[Badecker and Kuminiak, 2007] Badecker, W. and Kuminiak, F. (2007). 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. 9, 32, 54

[Bekinschtein et al., 2009a] Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., and Naccache, L. (2009a). Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences*, 106(5):1672–1677. 15

[Bekinschtein et al., 2009b] Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., and Naccache, L. (2009b). Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences*, 106(5):1672–1677. 15, 18, 19, 31, 83

[Berwick et al., 2011] Berwick, R. C., Pietroski, P., Yankama, B., and Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, 35(7):1207–1242. 4

[Berwick and Weinberg, 1986] Berwick, R. C. and Weinberg, A. S. (1986). *The grammatical basis of linguistic performance: Language use and acquisition*. MIT press. 3

[Bock et al., 2012] Bock, K., Carreiras, M., and Meseguer, E. (2012). Number meaning and number grammar in english and spanish. *Journal of Memory and Language*, 66(1):17–37. 28

[Bock and Cutting, 1992] Bock, K. and Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of memory and language*, 31(1):99–127. 8

[Bock and Eberhard, 1993] Bock, K. and Eberhard, K. M. (1993). Meaning, sound and syntax in english number agreement. *Language and Cognitive Processes*, 8(1):57–99. 8

[Bock and Miller, 1991] Bock, K. and Miller, C. A. (1991). Broken agreement. *Cognitive psychology*, 23(1):45–93. 8, 18, 28, 70, 73, 92

[Bornkessel and Schlesewsky, 2006] Bornkessel, I. and Schlesewsky, M. (2006). The extended argument dependency model: a neurocognitive approach to sentence comprehension across languages. *Psychological review*, 113(4):787. 85

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*. 73

[Burling, 2007] Burling, R. (2007). *The talking ape: How language evolved*, volume 5. Oxford University Press on Demand. 1

[Bybee, 2002] Bybee, I. (2002). Sequentiality as the basis. *The evolution of language out of pre-language*, 53:109. 4, 5

[Caffarra et al., 2019] Caffarra, S., Mendoza, M., and Davidson, D. (2019). Is the lan effect in morphosyntactic processing an erp artifact? *Brain and language*, 191:9–16. 8

[Caucheteux et al., 2021] Caucheteux, C., Gramfort, A., and King, J.-R. (2021). Long-range and hierarchical language predictions in brains and algorithms. 7

[Chen et al., 2007] Chen, L., Shu, H., Liu, Y., Zhao, J., and Li, P. (2007). Erp signatures of subject–verb agreement in l2 learning. *Bilingualism: Language and Cognition*, 10(2):161–174. 11

[Chomsky, 1957] Chomsky, N. (1957). *Syntactic structures*. Mouton publishers. Accepted: 2019-04-10T06:25:21Z Journal Abbreviation: Janua linguarum. 2, 36, 70, 83, 86

[Chomsky, 2009] Chomsky, N. (2009). *Syntactic structures*. De Gruyter Mouton. 2, 49, 86

[Chomsky, 2014a] Chomsky, N. (2014a). Minimal recursion: exploring the prospects. In *Recursion: Complexity in cognition*, pages 1–15. Springer. 2, 49

[Chomsky, 2014b] Chomsky, N. (2014b). *The minimalist program*. MIT press. 2, 53

[Chomsky, 2015] Chomsky, N. (2015). *What kind of creatures are we?* Columbia University Press. 4, 5

[Christiansen and Chater, 2015] Christiansen, M. H. and Chater, N. (2015). The language faculty that wasn't: A usage-based account of natural language recursion. *Frontiers in Psychology*, 6:1182. 5, 83

[Cinque and Rizzi, 2010] Cinque, G. and Rizzi, L. (2010). The cartography of syntactic structures. *Oxford Handbook of linguistic analysis*. 36

[Clark, 2013] Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204. 11

[Coopmans et al., 2021] Coopmans, C. W., De Hoop, H., Kaushik, K., Hagoort, P., and Martin, A. E. (2021). Structure-(in) dependent interpretation of phrases in humans and lstms. In *The Society for Computation in Linguistics (SCiL 2021)*, pages 459–463. 3, 52, 70

[Coulson et al., 1998] Coulson, S., King, J. W., and Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and cognitive processes*, 13(1):21–58. 8

[Dehaene and King, 2016] Dehaene, S. and King, J.-R. (2016). Decoding the dynamics of conscious perception: The temporal generalization method. *Micro-, meso-and macro-dynamics of the brain*, pages 85–97. 16, 24, 31, 41, 49, 94

[Dehaene et al., 2015] Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., and Pallier, C. (2015). The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron*, 88(1):2–19. 14, 15, 17, 19, 31, 36, 48, 70, 83, 85, 86

[DeLong et al., 2005] DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, 8(8):1117–1121. 11

[Dillon et al., 2013] Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2):85–103. 9, 32, 54

[Ding et al., 2017] Ding, N., Melloni, L., Tian, X., and Poeppel, D. (2017). Rule-based and word-level statistics-based processing of language: insights from neuroscience. *Language, cognition and neuroscience*, 32(5):570–575. 1

[Ding et al., 2015] Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2015). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19(1):158–164. 1, 70

[Duncan, 2010] Duncan, J. (2010). The multiple-demand (md) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences*, 14(4):172–179. 12, 53

[Eberhard, 1997] Eberhard, K. M. (1997). The marked effect of number on subject–verb agreement. *Journal of Memory and language*, 36(2):147–164. 28

[Eberhard et al., 2005] Eberhard, K. M., Cutting, J. C., and Bock, K. (2005). Making syntax of sense: number agreement in sentence production. *Psychological review*, 112(3):531. 9, 10, 54

[El Karoui et al., 2015] El Karoui, I., King, J.-R., Sitt, J., Meyniel, F., Van Gaal, S., Hasboun, D., Adam, C., Navarro, V., Baulac, M., Dehaene, S., et al. (2015). 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. 18

[Elman, 1990] Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211. 5

[Everaert et al., 2015] Everaert, M. B., Huybregts, M. A., Chomsky, N., Berwick, R. C., and Bolhuis, J. J. (2015). Structures, not strings: linguistics as part of the cognitive sciences. *Trends in cognitive sciences*, 19(12):729–743. 3

[Fayol et al., 1994] Fayol, M., Largy, P., and Lemaire, P. (1994). Cognitive overload and orthographic errors: When cognitive overload enhances subject–verb agreement errors. a study in french written language. *The Quarterly Journal of Experimental Psychology*, 47(2):437–464. 9

[Fedorenko et al., 2011] Fedorenko, E., Behr, M. K., and Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108(39):16428–16433. 12, 53

[Fedorenko et al., 2012] Fedorenko, E., Duncan, J., and Kanwisher, N. (2012). Language-selective and domain-general regions lie side by side within broca's area. *Current Biology*, 22(21):2059–2062. 12, 53

[Ferreira and Qiu, 2021] Ferreira, F. and Qiu, Z. (2021). Predicting syntactic structure. *Brain Research*, page 147632. 13

[Finlayson et al., 2021] Finlayson, M., Mueller, A., Gehrmann, S., Shieber, S., Linzen, T., and Belinkov, Y. (2021). Causal analysis of syntactic agreement mechanisms in neural language models. *arXiv preprint arXiv:2106.06087*. 70

[Fitch, 2014] Fitch, W. T. (2014). Toward a computational framework for cognitive biology: unifying approaches from cognitive neuroscience and comparative cognition. *Physics of life reviews*, 11(3):329–364. 14

[Fitch and Hauser, 2004] Fitch, W. T. and Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science*, 303(5656):377–380. 2

[Franck et al., 2007] Franck, J., Frauenfelder, U. H., and Rizzi, L. (2007). A syntactic analysis of interference in subject–verb agreement. *MIT working papers in linguistics*, (53):173–190. 19, 36

[Franck et al., 2006] Franck, J., Lassi, G., Frauenfelder, U. H., and Rizzi, L. (2006). Agreement and movement: A syntactic analysis of attraction. *Cognition*, 101(1):173–216. 36

[Franck et al., 2010] Franck, J., Soare, G., Frauenfelder, U. H., and Rizzi, L. (2010). Object interference in subject–verb agreement: The role of intermediate traces of movement. *Journal of Memory and Language*, 62(2):166–182. 36

[Franck et al., 2002] Franck, J., Vigliocco, G., and Nicol, J. (2002). Subject-verb agreement errors in french and english: The role of syntactic hierarchy. *Language and cognitive processes*, 17(4):371–404. 9, 28, 54

[Frank and Bod, 2011] Frank, S. L. and Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological science*, 22(6):829–834. 5, 13, 49, 70, 83

[Frank et al., 2012] Frank, S. L., Bod, R., and Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, 279(1747):4522–4531. 5, 13, 49

[Frank and Christiansen, 2018a] Frank, S. L. and Christiansen, M. H. (2018a). Hierarchical and sequential processing of language. *Language, Cognition and Neuroscience*, 33(9):1213–1218. 36

[Frank and Christiansen, 2018b] Frank, S. L. and Christiansen, M. H. (2018b). Hierarchical and sequential processing of language: A response to: Ding, melloni, tian, and poeppel (2017). rule-based and word-level statistics-based processing of language: insights from neuroscience. language, cognition and neuroscience. *Language, Cognition and Neuroscience*, 33(9):1213–1218. 5, 13, 49

[Frazier and Fodor, 1978] Frazier, L. and Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 6(4):291–325. 85

[Friederici, 2002] Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in cognitive sciences*, 6(2):78–84. 85

[Friederici, 2011] Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392. 3

[Friederici, 2017] Friederici, A. D. (2017). *Language in our brain: The origins of a uniquely human capacity*. MIT Press. 8

[Friederici et al., 2006a] Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., and Anwander, A. (2006a). The brain differentiates human and non-human grammars: Functional localization and structural connectivity. *Proceedings of the National Academy of Sciences*, 103(7):2458–2463. 2

[Friederici et al., 2006b] Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., and Anwander, A. (2006b). The brain differentiates human and non-human grammars: functional localization and structural connectivity. *Proceedings of the National Academy of Sciences*, 103(7):2458–2463. 2

[Friederici et al., 1996] Friederici, A. D., Hahne, A., and Mecklinger, A. (1996). Temporal structure of syntactic parsing: early and late event-related brain potential effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(5):1219. 8

[Friederici and Kotz, 2003] Friederici, A. D. and Kotz, S. A. (2003). The brain basis of syntactic processes: functional imaging and lesion studies. *Neuroimage*, 20:S8–S17. 8

[Friederici et al., 1999] Friederici, A. D., Steinhauer, K., and Frisch, S. (1999). Lexical integration: Sequential effects of syntactic and semantic information. *Memory & cognition*, 27(3):438–453. 22

[Friederici and Weissenborn, 2007] Friederici, A. D. and Weissenborn, J. (2007). Mapping sentence form onto meaning: The syntax–semantic interface. *Brain research*, 1146:50–58. 85

[Friston, 2005] Friston, K. (2005). A theory of cortical responses. 360(1456):815–836. 11, 21, 36

[Friston, 2010] Friston, K. (2010). The free-energy principle: a unified brain theory? 11(2):127–138. 11

[Friston et al., 2021] Friston, K., Moran, R. J., Nagai, Y., Taniguchi, T., Gomi, H., and Tenenbaum, J. (2021). World model learning and inference. 144:573–590. 21, 36, 70, 86

[Gibson et al., 2013] Gibson, E., Bergen, L., and Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056. 36

[Goodkind and Bicknell, 2021] Goodkind, A. and Bicknell, K. (2021). Local word statistics affect reading times independently of surprisal. *arXiv preprint arXiv:2103.04469*. 70, 76

[Gordon et al., 2001] Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *Journal of experimental psychology: learning, memory, and cognition*, 27(6):1411. 10

[Gordon et al., 2002] Gordon, P. C., Hendrick, R., and Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychological science*, 13(5):425–430. 10

[Gramfort et al., 2013] Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al. (2013). Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 7:267. 23, 24, 41

[Hagoort, 2003] Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: Erp effects of combining syntactic and semantic violations. *Journal of cognitive neuroscience*, 15(6):883–899. 22

[Hagoort et al., 1993] Hagoort, P., Brown, C., and Groothusen, J. (1993). The syntactic positive shift (sps) as an erp measure of syntactic processing. *Language and cognitive processes*, 8(4):439–483. 8

[Hagoort et al., 2003] Hagoort, P., Wassenaar, M., and Brown, C. (2003). Real-time semantic compensation in patients with agrammatic comprehension: Electrophysiological evidence for multiple-route plasticity. *Proceedings of the National Academy of Sciences*, 100(7):4340–4345. 8

[Hahne and Friederici, 1999] Hahne, A. and Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of cognitive neuroscience*, 11(2):194–205. 8

[Hale, 2016] Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412. 5

[Hammerly et al., 2019] Hammerly, C., Staub, A., and Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive psychology*, 110:70–104. 9, 10, 70, 85, 92

[Haskell and MacDonald, 2005] Haskell, T. R. and MacDonald, M. C. (2005). Constituent structure and linear order in language production: evidence from subject-verb agreement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):891. 1, 5, 70

[Haskell et al., 2010] Haskell, T. R., Thornton, R., and MacDonald, M. C. (2010). Experience and grammatical agreement: Statistical learning shapes number agreement production. *Cognition*, 114(2):151–164. 5

[Haspelmath, 2006] Haspelmath, M. (2006). Against markedness (and what to replace it with). *Journal of linguistics*, 42(1):25–70. 8

[Hauser et al., 2002] Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579. 2, 3, 49, 70, 83

[Heilbron et al., 2021] Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., and de Lange, F. P. (2021). A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv*, pages 2020–12. 13, 14, 36, 53

[Herrmann et al., 2011] Herrmann, B., Maess, B., and Friederici, A. D. (2011). Violation of syntax and prosody—disentangling their contributions to the early left anterior negativity (elan). *Neuroscience letters*, 490(2):116–120. 8

[Huettig and Guerra, 2015] Huettig, F. and Guerra, E. (2015). Testing the limits of prediction in language processing: Prediction occurs but far from always. In *the 21st Annual Conference on Architectures and Mechanisms for Language Processing (AMLaP 2015)*. 12

[Huettig and Mani, 2016] Huettig, F. and Mani, N. (2016). Is prediction necessary to understand language? probably not. *Language, Cognition and Neuroscience*, 31(1):19–31. 12

[Huettig et al., 2011] Huettig, F., Rommers, J., and Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. 137(2):151–171. 11

[Jackendoff, 1972] Jackendoff, R. S. (1972). Semantic interpretation in generative grammar. 3

[Jackendoff, 2008] Jackendoff, R. S. (2008). *Patterns in the mind: Language and human nature*. Basic Books. 1

[Jäger et al., 2017] Jäger, L. A., Engelmann, F., and Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and bayesian meta-analysis. *Journal of Memory and Language*, 94:316–339. 10

[Jas et al., 2018] Jas, M., Larson, E., Engemann, D. A., Leppäkangas, J., Taulu, S., Hämäläinen, M., and Gramfort, A. (2018). A reproducible meg/eeg group study with the mne software: recommendations, quality assessments, and good practices. *Frontiers in neuroscience*, 12:530. 23, 41

[Jonides et al., 2008] Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., and Moore, K. S. (2008). The mind and brain of short-term memory. *Annu. Rev. Psychol.*, 59:193–224. 9, 85

[Jumelet et al., 2019] Jumelet, J., Zuidema, W., and Hupkes, D. (2019). Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. *arXiv preprint arXiv:1909.08975*. 70

[Kaan, 2002] Kaan, E. (2002). Investigating the effects of distance and number interference in processing subject-verb dependencies: An erp study. *Journal of Psycholinguistic Research*, 31(2):165–193. 8, 11

[Kaan et al., 2000] Kaan, E., Harris, A., Gibson, E., and Holcomb, P. (2000). The p600 as an index of syntactic integration difficulty. *Language and cognitive processes*, 15(2):159–201. 8

[Keller and Mrsic-Flogel, 2018] Keller, G. B. and Mrsic-Flogel, T. D. (2018). Predictive processing: a canonical cortical computation. *Neuron*, 100(2):424–435. 11

[King and Dehaene, 2014] King, J.-R. and Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, 18(4):203–210. 15, 16, 18, 19, 24, 32, 41, 49, 94

[King et al., 2018] King, J.-R., Gwilliams, L., Holdgraf, C., Sassenhagen, J., Barachant, A., Engemann, D., Larson, E., and Gramfort, A. (2018). Encoding and decoding neuronal dynamics: Methodological framework to uncover the algorithms of cognition. 32

[Kratzer and Heim, 1998] Kratzer, A. and Heim, I. (1998). *Semantics in generative grammar*, volume 1185. Blackwell Oxford. 3

[Kuperberg and Jaeger, 2016] Kuperberg, G. R. and Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, 31(1):32–59. 21

[Kutas and Hillyard, 1983] Kutas, M. and Hillyard, S. A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory & cognition*, 11(5):539–550. 8

[Kutas and Hillyard, 1984] Kutas, M. and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163. 11

[Lago et al., 2015] Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., and Phillips, C. (2015). Agreement processes in spanish comprehension. *Journal of Memory and Language*, 82:133–149. 28, 32

[Lakretz et al., 2021a] Lakretz, Y., Desbordes, T., Hupkes, D., and Dehaene, S. (2021a). Causal transformers perform below chance on recursive nested constructions, unlike humans. 7

[Lakretz et al., 2021b] Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., and Dehaene, S. (2021b). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699. Special Issue in Honour of Jacques Mehler, Cognition's founding editor. 6, 7

[Lakretz et al., 2021c] Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., and Dehaene, S. (2021c). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699. 18

[Lakretz et al., 2019a] Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., and Baroni, M. (2019a). The emergence of number and syntax units in lstm language models. 5, 6, 18, 31, 32, 85, 87

[Lakretz et al., 2019b] Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., and Baroni, M. (2019b). The emergence of number and syntax units in LSTM language models. *CoRR*, abs/1903.07435. 70

[Lau et al., 2006] Lau, E., Stroud, C., Plesch, S., and Phillips, C. (2006). The role of structural prediction in rapid syntactic analysis. *Brain and language*, 98(1):74–88. 13

[Levy, 2008] Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177. 36

[Lewis and Vasishth, 2005] Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, 29(3):375–419. 9, 10, 32, 54, 85

[Lidz et al., 2003] Lidz, J., Waxman, S., and Freedman, J. (2003). What infants know about syntax but couldn't have learned: experimental evidence for syntactic structure at 18 months. *Cognition*, 89(3):295–303. 3

[Linzen and Baroni, 2021] Linzen, T. and Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212. 5

[Linzen et al., 2016a] Linzen, T., Dupoux, E., and Goldberg, Y. (2016a). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535. 5, 6

[Linzen et al., 2016b] Linzen, T., Dupoux, E., and Goldberg, Y. (2016b). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535. 70

[Lorimor et al., 2008] Lorimor, H., Bock, K., Zalkind, E., Sheyman, A., and Beard, R. (2008). Agreement and attraction in russian. *Language and cognitive processes*, 23(6):769–799. 28

[Makuuchi et al., 2009] Makuuchi, M., Bahlmann, J., Anwander, A., and Friederici, A. D. (2009). Segregating the core computational faculty of human language from working memory. *Proceedings of the National Academy of Sciences*, 106(20):8362–8367. 2

[Mani et al., 2012] Mani, N., Durrant, S., and Floccia, C. (2012). Activation of phonological and semantic codes in toddlers. *Journal of Memory and Language*, 66(4):612–622. 12

[Maris and Oostenveld, 2007] Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190. xxvii, 24, 26, 31, 41, 47, 48, 63, 94

[Martin, 2020] Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience*, 32(8):1407–1427. 3

[Martin and McElree, 2008] Martin, A. E. and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language*, 58(3):879–906. 9, 32, 54

[McElree et al., 2003] McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of memory and language*, 48(1):67–91. 9, 32, 54, 85

[Molinaro et al., 2015] Molinaro, N., Barber, H. A., Caffarra, S., and Carreiras, M. (2015). On the left anterior negativity (lan): The case of morphosyntactic agreement. *Cortex*, 66(156-159). 8

[Molinaro et al., 2011a] Molinaro, N., Barber, H. A., and Carreiras, M. (2011a). Grammatical agreement processing in reading: Erp findings and future directions. *cortex*, 47(8):908–930. 8, 36, 70, 87

[Molinaro et al., 2011b] Molinaro, N., Vespignani, F., Zamparelli, R., and Job, R. (2011b). Why brother and sister are not just siblings: Repair processes in agreement computation. *Journal of Memory and Language*, 64(3):211–232. 8

[Napoles et al., 2017] Napoles, C., Sakaguchi, K., and Tetreault, J. (2017). Jfleg: A fluency corpus and benchmark for grammatical error correction. 73

[Nave et al., 2020] Nave, K., Deane, G., Miller, M., and Clark, A. (2020). Wilding the predictive brain. *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(6):e1542. 11

[Nelson et al., 2017] Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., et al. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114(18):E3669–E3678. 3, 52, 70

[Norvig, 2017] Norvig, P. (2017). On chomsky and the two cultures of statistical learning. In *Berechenbarkeit der Welt?*, pages 61–83. Springer. 5

[Osterhout and Holcomb, 1992] Osterhout, L. and Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, 31(6):785–806. 8, 21

[Osterhout and Mobley, 1995] Osterhout, L. and Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and language*, 34(6):739–773. 8, 21

[Paape et al., 2021] Paape, D., Avetisyan, S., Lago, S., and Vasishth, S. (2021). Modeling misretrieval and feature substitution in agreement attraction: A computational evaluation. *Cognitive Science*, 45(8):e13019. 70

[Pallier et al., 2011] Pallier, C., Devauchelle, A.-D., and Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527. 3, 52, 70

[Partee, 1975] Partee, B. (1975). Montague grammar and transformational grammar. *Linguistic inquiry*, pages 203–300. 3

[Pearlmutter et al., 1999] Pearlmutter, N. J., Garnsey, S. M., and Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and language*, 41(3):427–456. 72

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830. 24, 41

[Pegado et al., 2010] Pegado, F., Bekinschtein, T., Chausson, N., Dehaene, S., Cohen, L., and Naccache, L. (2010). Probing the lifetimes of auditory novelty detection processes. *Neuropsychologia*, 48(10):3145–3154. 15, 19

[Perrin et al., 1989] Perrin, F., Pernier, J., Bertrand, O., and Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology*, 72(2):184–187. 23, 41

[Phillips et al., 2011] Phillips, C., Wagers, M. W., and Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. *Experiments at the Interfaces*, 37:147–180. 8, 73

[Pinker, 1998] Pinker, S. (1998). Words and rules. *Lingua*, 106(1-4):219–242. 3

[Raffel et al., 2019] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*. 73

[Ristic et al., 2021] Ristic, B., Mancini, S., Molinaro, N., and Staub, A. (2021). Maintenance cost in the processing of subject–verb dependencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 10

[Rizzi, 2004] Rizzi, L. (2004). On the cartography of syntactic structures. *The Structure of CP and IP*, pages 3–15. 36, 49, 70, 73, 83, 86

[Rossi et al., 2005] Rossi, S., Gugler, M. F., Hahne, A., and Friederici, A. D. (2005). When word category information encounters morphosyntax: An erp study. *Neuroscience Letters*, 384(3):228–233. 8

[Santesteban et al., 2017] Santesteban, M., Zawiszewski, A., Erdocia, K., and Laka, I. (2017). On the nature of clitics and their sensitivity to number attraction effects. *Frontiers in psychology*, 8:1470. 11

[Schlueter et al., 2019] Schlueter, Z., Parker, D., and Lau, E. (2019). Error-driven retrieval in agreement attraction rarely leads to misinterpretation. *Frontiers in psychology*, 10:1002. 54, 85

[Severens et al., 2008] Severens, E., Jansma, B. M., and Hartsuiker, R. J. (2008). Morphophonological influences on the comprehension of subject–verb agreement: An erp study. *Brain Research*, 1228:135–144. 11

[Shain, 2019] Shain, C. (2019). A large-scale study of the effects of word frequency and predictability in naturalistic reading. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics. 76

[Shain et al., 2020] Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., and Fedorenko, E. (2020). fmri reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307. 12, 13, 53

[Shain and Schuler, 2021] Shain, C. and Schuler, W. (2021). Continuous-time deconvolutional regression for psycholinguistic modeling. *Cognition*, 215:104735. 12, 13

[Shen et al., 2013] Shen, E. Y., Staub, A., and Sanders, L. D. (2013). Event-related brain potential evidence that local nouns affect subject–verb agreement processing. *Language and Cognitive Processes*, 28(4):498–524. 11, 70

[Shi et al., 2020] Shi, R., Legrand, C., and Brandenberger, A. (2020). Toddlers track hierarchical structure dependence. *Language Acquisition*, 27(4):397–409. 3, 52, 70

[Sinha et al., 2021] Sinha, K., Jia, R., Hupkes, D., Pineau, J., Williams, A., and Kiela, D. (2021). Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*. 70

[Sonnenschein, 1925] Sonnenschein, E. (1925). The philosophy of grammar-the philosophy of grammar. by otto jespersen. 8vo. pp. 359. london: George allen and unwin, 1924. cloth, 12s. 6d. net. *The Classical Review*, 39(1-2):38–40. 9, 15

[Staub, 2009] Staub, A. (2009). On the interpretation of the number attraction effect: Response time evidence. *Journal of memory and language*, 60(2):308–327. 9, 54

[Staub, 2010] Staub, A. (2010). Response time distributional evidence for distinct varieties of number attraction. *Cognition*, 114(3):447–454. 9, 54

[Steinhauer and Drury, 2012] Steinhauer, K. and Drury, J. E. (2012). On the early left-anterior negativity (elan) in syntax studies. *Brain and language*, 120(2):135–162. 8

[Stoops and Christianson, 2017] Stoops, A. and Christianson, K. (2017). Parafoveal processing of inflectional morphology on russian nouns. *Journal of Cognitive Psychology*, 29(6):653–669. 32, 85

[Strauss et al., 2015] Strauss, M., Sitt, J. D., King, J.-R., Elbaz, M., Azizi, L., Buiatti, M., Naccache, L., Van Wassenhove, V., and Dehaene, S. (2015). Disruption of hierarchical predictive coding during sleep. *Proceedings of the National Academy of Sciences*, 112(11):E1353–E1362. 15, 18, 19

[Tanner and Bulkes, 2015] Tanner, D. and Bulkes, N. Z. (2015). Cues, quantification, and agreement in language comprehension. *Psychonomic bulletin & review*, 22(6):1753–1763. 10, 11

[Tanner et al., 2018] Tanner, D., Goldshtein, M., and Weissman, B. (2018). Individual differences in the real-time neural dynamics of language comprehension. In *Psychology of learning and motivation*, volume 68, pages 299–335. Elsevier. 10, 16

[Tanner et al., 2017] Tanner, D., Grey, S., and van Hell, J. G. (2017). Dissociating retrieval interference and reanalysis in the p600 during sentence comprehension. *Psychophysiology*, 54(2):248–259. 11, 53, 54, 86

[Tanner et al., 2015] Tanner, D., Morgan-Short, K., and Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in erp studies of language and cognition. *Psychophysiology*, 52(8):997–1009. 10, 16

[Tanner et al., 2014] Tanner, D., Nicol, J., and Brehm, L. (2014). The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of memory and language*, 76:195–215. 10, 11, 70

[Tanner and Van Hell, 2014] Tanner, D. and Van Hell, J. G. (2014). Erps reveal individual differences in morphosyntactic processing. *Neuropsychologia*, 56:289–301. 8, 85

[Taulu et al., 2004] Taulu, S., Kajola, M., and Simola, J. (2004). Suppression of interference and artifacts by the signal space separation method. *Brain topography*, 16(4):269–275. 23, 41

[Uddén et al., 2020] Uddén, J., de Jesus Dias Martins, M., Zuidema, W., and Tecumseh Fitch, W. (2020). Hierarchical structure in sequence processing: How to measure it and determine its neural implementation. *Topics in cognitive science*, 12(3):910–924. 4, 49, 70

[Uhrig et al., 2014] Uhrig, L., Dehaene, S., and Jarraya, B. (2014). A hierarchy of responses to auditory regularities in the macaque brain. *Journal of Neuroscience*, 34(4):1127–1132. 15, 19

[Vagharchakian et al., 2012] Vagharchakian, L., Dehaene-Lambertz, G., Pallier, C., and Dehaene, S. (2012). A temporal bottleneck in the language comprehension network. *Journal of Neuroscience*, 32(26):9089–9102. 37

[Van Dyke and Johns, 2012] Van Dyke, J. A. and Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Language and linguistics compass*, 6(4):193–211. 9, 32, 54, 85

[Vasishth et al., 2017] Vasishth, S., Jäger, L. A., and Nicenboim, B. (2017). Feature overwriting as a finite mixture process: Evidence from comprehension data. *arXiv preprint arXiv:1703.04081*. 10

[Vigliocco and Nicol, 1998] Vigliocco, G. and Nicol, J. (1998). Separating hierarchical relations and word order in language production: Is proximity concord syntactic or linear? *Cognition*, 68(1):B13–B29. 9, 49, 54, 70

[Villata et al., 2018] Villata, S., Tabor, W., and Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. *Frontiers in psychology*, 9:2. 10

[Wagers et al., 2009] Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of memory and language*, 61(2):206–237. x, xxiii, 8, 9, 25, 28, 32, 43, 54, 56, 73, 77, 85, 86, 90

[Willer Gold et al., 2017] Willer Gold, J., Arsenijević, B., Batinić, M., Becker, M., Čordalija, N., Kresić, M., Leko, N., Marušič, F. L., Milićev, T., Milićević, N., et al. (2017). When linearity prevails over hierarchy in syntax. pnas. 1, 70

[Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Huggingface's transformers: State-of-the-art natural language processing. 72

[Yujian and Bo, 2007] Yujian, L. and Bo, L. (2007). A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095. 40

[Zaccarella and Friederici, 2015] Zaccarella, E. and Friederici, A. D. (2015). Merge in the human brain: A sub-region based functional investigation in the left pars opercularis. *Frontiers in psychology*, 6:1818. 3

[Zandvoort, 1961] Zandvoort, R. (1961). Varia syntactica. language and society: Essays presented to arthur m. jensen on his seventieth birthday. 8