A search for the neural bases of compositionality

Théo Desbordes

FACEBOOK AI

École Doctorale Cerveau Cognition Comportement

À LA RECHERCHE DES BASES NEURALES DE LA COMPOSITIONNALITÉ

Par Théo DESBORDES

Thèse de doctorat de neurosciences cognitives

dirigée par Stanislas DEHAENE et co-dirigée par Jean-Rémi KING

Soutenance publique effectuée le 13 décembre 2022 à 13h12

Devant un jury composé de :

Lucia MELLONI	Max Planck Institute for Empirical Aesthetics	Rapporteuse
Laura GWILLIAMS	University of California, San Francisco	Rapporteuse
Christophe PALLIER	Neurospin	Président du jury
Alexandre GRAMFORT	INRIA	Examinateur
Stanislas DEHAENE	Collège de France	Examinateur
Jean-Rémi KING	École normale Supérieure & Meta	Examinateur

École Doctorale Cerveau Cognition Comportement

A SEARCH FOR THE NEURAL BASES OF COMPOSITIONALITY

Théo DESBORDES

Dissertation submitted for the degree of Doctor of Philosophy in Cognitive Neurosciences

Supervised by Stanislas DEHAENE et co-supervised by Jean-Rémi KING

Public defense took place on the 13th of December 2022

Jury :

Lucia MELLONI	Max Planck Institute for Empirical Aesthetics	Rapporteur
Laura GWILLIAMS	University of California, San Francisco	Rapporteur
Christophe PALLIER	Neurospin	Jury's president
Alexandre GRAMFORT	INRIA	Examiner
Stanislas DEHAENE	Collège de France	Examiner
Jean-Rémi KING	École normale Supérieure & Meta	Examiner

Abstract

Wilhelm von Humboldt famously said that language makes “infinite use of finite means”. Concretely, humans are able to combine words following a fixed set of grammatical rules, yielding a limitless reservoir of meanings. This is what I mean by compositionality, and what this thesis has aimed to characterize. Using a combination of tools from linguistics, natural language processing and neuroimaging, I sought to describe how this impressive feat is implemented in the human brain.

In the first part, I present a study that examines the rise of compositional representations. We isolate semantic processes by comparing normal sentences with ones made of meaningless pseudowords, *aka* Jabberwocky. Using joint magnetoencephalography and intracranial electroencephalography recordings, we show that the intrinsic dimensionality of the neural signals grows over time, and more so for normal sentences than Jabberwocky, portraying the progressive recruitment of neurons in the semantic representation. Furthermore, by means of multivariate decoding, we demonstrate that the dynamics of neural signals follow theoretically driven patterns, especially ramping and sentence-final signatures. In addition, we take advantage of the fine spatial resolution of intracranial recordings to quantify the participation of various brain regions in each of these steps and identify a chief role of the prefrontal cortex in compositional processes. Crucially, these signatures are present in state-of-the-art neural language models, but absent in untrained models, suggesting that learning language is associated with a predictable shaping of the neural vector space. Overall, we show that the neural representations of sentences grow with each additional meaning that can be added to the existing semantic manifold.

The second study takes a finer experimental approach by focusing on the cortical representation of phrases composed of a small number of nouns and adjectives, in a working memory task with distinct sentence encoding, delay, and picture comparison stages. Using magnetoencephalography recordings and multivariate decoding, we collect brain responses to words in isolation and within increasingly longer phrases, and use these data to investigate the organization and temporal evolution of compositional representations. During the encoding phase, a cascade of activations follows each new

word, and crucially, the representation of individual words is partially sustained until it can be coherently integrated into a phrase, at which point it fades away. Later, during the delay period, in which the subjects keep in mind the sentence to match it to a subsequent image, neural activity reflects the complexity of the sentence, as quantified by the number of different words it comprises. Finally, when the compositional representation has to be read-out, the speed of this mechanism is also modulated by complexity, as well as by the syntactic depth of the query: surface properties are detected faster than syntactically deeper ones. These findings suggest that the compositional word representations are compressed in working memory and require task-specific decompression to be accessed.

Taken together, these findings shed new light on the nature of compositional representations in the human brain. Both studies point towards the idea that semantic representations are encoded in distributed vector spaces, perhaps similar to artificial neural language models and vector-symbolic architectures. We provide the first steps towards the characterization of these neural semantic spaces, their dimensionality and how they evolve over time.

Résumé

Wilhelm von Humboldt a déclaré que le langage « fait un usage infini de moyens finis ». En effet, les humains sont capables de combiner des mots en suivant un ensemble fixe de règles grammaticales, produisant ainsi un réservoir illimité de sens. C'est ce que j'entends par compositionnalité et ce que cette thèse s'est attachée à caractériser. En utilisant une combinaison d'outils venant de la linguistique, du traitement automatique du langage et de la neuroimagerie, j'ai cherché à décrire comment cet exploit est mis en œuvre dans le cerveau humain.

Dans la première partie, je présente une étude qui examine l'émergence des représentations compositionnelles. Nous isolons les processus sémantiques en comparant des phrases normales avec celles faites de pseudo-mots dépourvus de sens, ou « Jabberwocky ». Dans une rare combinaison d'enregistrements de magnétoencéphalographie et d'électroencéphalographie intracrânienne, nous montrons que la dimensionnalité intrinsèque des signaux neuronaux croît avec le temps, et plus encore pour les phrases normales que le Jabberwocky, dépeignant le recrutement progressif des neurones dans la représentation compositionnelle. De plus, au moyen de décodage multivarié, nous démontrons que la dynamique des représentations suit des schémas théoriques, en particulier de rampe et de fin de phrase. De plus, nous profitons de la résolution spatiale fine des enregistrements intracrâniens pour quantifier la participation de différentes régions du cerveau à chacune de ces étapes et identifier notamment le rôle principal du cortex frontal dans les processus compositionnels. Crucialement, ces signatures étaient présentes dans des modèles de langage de pointe, mais absentes dans les modèles non entraînés, ce qui suggère que l'apprentissage du langage est associé à un remodelage prévisible de l'espace vectoriel neuronal. Finalement, nous montrons que les représentations neuronales croissent avec chaque signification supplémentaire qui peut être ajoutée à la variété sémantique existante.

La deuxième étude adopte une approche expérimentale plus minutieuse en se concentrant sur les représentations corticales de phrases composées d’un petit nombre de noms et d'adjectifs, dans une tache de mémoire de travail avec des étapes distinctes

d’encodage, de délai et de comparaison avec une image. À l'aide d'enregistrements magnétoencéphalographiques et du décodage multivarié, nous collectons les réponses cérébrales à des mots isolés ainsi que dans des phrases de plus en plus longues. Pendant la phase d'encodage, une cascade d'activations suit chaque nouveau mot et, surtout, la représentation des mots individuels est partiellement maintenue jusqu'à ce qu'elle puisse être intégrée de manière cohérente dans une phrase, après quoi elle s'estompe. Ensuite, pendant une période de délai, au cours de laquelle les sujets devaient retenir la phrase pour la comparer à une image ultérieure, l'activité neuronale reflète la complexité de la phrase, quantifiée par le nombre de mots différents qu'elle contient. Enfin, lorsque la représentation compositionnelle doit être lue, la vitesse de ce mécanisme est également modulée par la complexité, ainsi que par la profondeur syntaxique de la requête. Ces résultats suggèrent que les représentations compositionnelles sont compressées dans la mémoire de travail et nécessitent une décompression spécifique pour être accédées.

Pris ensemble, ces résultats ouvrent une fenêtre sur la nature des représentations compositionnelles dans le cerveau humain. Les deux études pointent vers l'idée que les représentations sémantiques sont codées dans des espaces vectoriels distribués, peut-être semblables aux modèles de langage artificiel et aux architectures vectorielles symboliques. Nous proposons un premier pas vers la caractérisation de ces espaces sémantiques neuronaux, leur dimensionnalité et leur évolution dans le temps.

Remerciements

J'ai eu le privilège d'avoir été supervisé par Jean-Rémi King et Stanislas Dehaene. J'ai énormément appris à leurs côtés et je leur en suis grandement reconnaissant. Ils ont su me transmettre leur passion intense pour la science et m'ont poussé à donner le meilleur de moi-même. Jean-Rémi a su me garder motivé même aux moments les plus difficiles, par son enthousiasme et ses retours toujours constructifs. Stanislas m'a poussé plus loin que ce que je croyais possible. J'ai également eu la chance d'avoir un troisième encadrant, Yair Lakretz, qui m'a conseillé et aiguillé pendant la majeure partie de cette thèse. Merci à vous trois pour votre temps et vos conseils !

Je souhaite également remercier tou.t.e.s les fantastiques collègues avec lesquels j'ai eu l'occasion d'échanger et de collaborer : Mathias, Christos, Lorenzo, Fosca, Lucas, Charlotte, Pierre, Alexandre, Christophe, Caroline, Minye, Pauline, Christian, Valérie, et les collaborateurs marseillais que je n'ai pas encore eu l'occasion de rencontrer.

Je remercie Marco pour m'avoir introduit à FAIR et Maxime pour avoir été mon manager préféré. Je remercie également tout le personnel de Neurospin et en particulier les infirmières et les manipulatrices radio qui m'ont aidé lors des acquisitions. Merci à Meta de m'avoir payé pour faire la même recherche que j'aurais faite dans le public.

Un merci tout particulier à mes précédents mentors. Jacques Brocard, mon premier superviseur, qui a confirmé mon envie de faire des neurosciences. Sébastien Marti, tristement parti trop tôt, qui a éveillé ma passion de l'analyse des données neuronales. Enfin, Lyle Muller et Terry Sejnowsky qui m'ont initié à la modélisation computationnelle.

Je souhaite également remercier un certain nombre de professeurs ayant eu un impact particulier lors de mes études : Mr. Massliah, Yann Essautier, Nicolas Bergasse, François Berger, Michel Sève, Mariano Casado, Boris Barbour, Éric Michel, Yves Boubenec, Étienne Koechlin, Sophie Denève, Srdjan Ostojic, Emmanuel Dupoux, Shihab Shamma.

Toute ma gratitude à Vanna et Patricia, sans qui le fardeau administratif aurait été insupportable, ainsi que pour leur général awesomeness.

Je remercie les relecteurs de mes travaux, et tout particulièrement les membres du jury. Merci également aux dictionnaires de synonymes, notamment thesaurus.com, sans qui ce manuscrit serait beaucoup moins digeste.

Je n'aurais jamais tenu la distance sans l'immense soutien de mes amis et de ma famille (et belle famille !). Je remercie en particulier l'équipe des coguys : Anns, Camille, Charlie, Joffrey, Jules, Julie, Kévin, Luiza, Mehdi, Morgan, Morgane, Pierre (bonjour Pierre !) et les bio15 : Étienne, Bobby, Corentin, Thomas, … et tou.t.e.s les autres ! Merci à Raph, Chiara, Élise, Manon, Arnaud, Guillain, Jean, Gaspard, Julie, Marius, Alexis, Clémence, Alexandre, Eva, Vlad, et tout.e.s mes binômes de grimpe et de vie.

Merci aux musiciens qui réussissent toujours à me donner sourire et motivation, en particulier Gojira, The Ocean, Rivers of Nihil. Merci aussi aux communautés de memes « High impact memes for PhD fiends », « Reviewer 2 Must Be Stopped! », « Journal of Scientific Shitposting (JoSS) », « Science diagrams that look like shitposts », and « Shitposts that look like science diagrams ».

Enfin, merci à Roland et Bernadette, sans qui je ne serais pas là aujourd'hui. Merci à mes cousin.e.s, oncles et tantes, et Madelon en particulier.

Merci à Damien et Lucie.

Merci à la Docteure.

Merci à Carole.

à Marie,

Table of Contents

Table of Figures

Publications of the author

Publications included in the thesis

Chapter 1. Dimensionality and ramping: signatures of sentence integration in the dynamics of brains and deep language models (in press in the Journal of Neuroscience)

Chapter 2. Evidence for a compressed neural code in working memory during language composition (submitted to Neuron)

Other publications

- *The emergence of number and syntax units in LSTM language models. Lakretz, Y., Kruszewski, G., **Desbordes, T.**, Hupkes, D., Dehaene, S., & Baroni, M. (2019). Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 11–20. https://doi.org/10.18653/v1/N19-1002*
- *Can RNNs learn Recursive Nested Subject-Verb Agreements? Lakretz, Y.*, **Desbordes, T.***, King, J.-R., Crabbé, B., Oquab, M., & Dehaene, S. (2021). ArXiv:2101.02258 [Cs]. * shared first authorship. http://arxiv.org/abs/2101.02258*
- *Causal Transformers Perform Below Chance on Recursive Nested Constructions, Unlike Humans , Lakretz, Y., **Desbordes, T.**, Hupkes, D., & Dehaene, S. (2021). (arXiv:2110.07240). arXiv. http://arxiv.org/abs/2110.07240*
- *Can Transformers Process Recursive Nested Constructions, Like Humans? Lakretz, Y., **Desbordes, T.**, Hupkes, D., & Dehaene, S. (2022, October). In Proceedings of the 29th International Conference on Computational Linguistics (pp. 3226-3232). https://aclanthology.org/2022.coling-1.285/*
- *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … **Desbordes, T.**, … Wu, Z. (2022). (arXiv:2206.04615). arXiv. http://arxiv.org/abs/2206.04615*
- *Isolating non-structural effects in grammatical agreement, Zacharopoulos, N.-C., Sablé-Meyer, M., **Desbordes, T.** (2022, in prep).*

Introduction

"[…] un des plus grands avantages de l'homme au-dessus de tous les autres animaux, et qui est une des plus grandes preuves de la raison : c'est l'usage que nous en faisons pour signifier nos pensées, et cette invention merveilleuse de composer de vingt-cinq ou trente sons cette infinie variété de mots, qui, n'ayant rien de semblable en eux-mêmes à ce qui se passe dans notre esprit, ne laissent pas d'en découvrir aux autres tout le secret, et de faire entendre à ceux qui n'y peuvent pénétrer, tout ce que nous concevons, et tous les divers mouvements de notre âme."

"[…] one of the great advantages of human beings compared to other animals, and which is one of the most significant proofs of reason: that is, the method by which we are able to express our thoughts, the marvelous invention by which using twenty five or thirty sounds we can create the infinite variety of words, which having nothing themselves in common with what is passing in our minds nonetheless permit us to express all our secrets, and which allow us to understand what is not present to consciousness, in effect, everything that we can conceive and the most diverse movements of our soul."

Antoine Arnault and Claude Lancelot, Grammaire générale et raisonnée de Port Royal (Arnauld & Lancelot, Claude, 1660)

A. Motivations to study language processing

Human languages are marvelous. Starting with a small, fixed set of sounds or characters, we couple them to spawn many more words. Then, we are able to combine elements of this finite vocabulary with one-another to generate ever new meanings. Take this sentence for example: *"Mick Jagger and the seven drumming jellyfish are rocking the place at Pompei in 3009 AD"*. You (most likely) haven't heard it before, but knowing the individual words and being a master of the rules to combine them, you spontaneously reconstruct its signification in your mind. This is what we mean by compositionality and what I tried to address in this thesis.

Compositionality could be the cornerstone of human intelligence, and is certainly a necessary condition for complex thoughts and creativity. Indeed, meaning composition has been called the *'holy grail'* of cognitive science (R. Jackendoff, 2002) and it has

fascinated scholars for millennia (Pagin & Westerståhl, 2010). As such, the study of language allows a one-of-a-kind window into the human mind. That is the main motivation for the choice of my thesis subject. Subjacent to this overarching goal, there are many motivations, laid out below.

First and foremost, language is the medium of some of the most amazing achievements of the human race: stories, songs, encyclopedias, constitutions, theses... everything is articulated with language (pun intended). In essence, it can be viewed as an endowment for hierarchical constructions, with phonemes assembling into words, words into sentences, sentences into narratives. To such a degree that it allows us to express "*everything that we can conceive and the most diverse movements of our soul" (Arnauld & Lancelot, Claude, 1660)*.

Second, it is omnipresent in our day-to-day life. We use it all the time, on most occasions without any effort. Not only is it the most efficient form of interpersonal communication, but our own thoughts are also expressed with language (the "little voice" in our heads that most people have, though not everyone (Alderson-Day & Fernyhough, 2015; Heavey et al., 2019)). So much so that when things go wrong, incurs debilitating diseases, such as in aphasia.

Finally, only in humans is it so proficiently developed and as such it might help understand what makes us unique. It is likely that the ability to communicate thoughts with each other was critical in our development. But perhaps even more critical could be the ability to flexibly manipulate concepts, to combine them and create new associations. This creativity, brought about by compositionality, is the hallmark of language. We have no problem understanding combinations of words that we have never heard before, such as the ludicrous "*Mick Jagger*" sentence above. Even more impressive, the sentence doesn't even have to make sense, as long as it is well formed, as in "The schmilblick did a backwards strumpf". What happens in our brain when we read such a sentence?

Intertwined with compositionality is the ability to generalize. Children, when they learn to speak, clearly demonstrate this. A word learned in one context will be used in another context, another sentence. Sometimes wrongly so, resulting in over-

generalization, e.g., using “cooker” to describe a person who cooks after learning “driver”, “presenter”, etc

Most remarkable in the linguistic domain is the ability to recursively combine elements. Basically, any constituents can be embedded to construct a larger, richer arrangement, as in e.g. “*The mother of the mayor that went to school with the doctor of the…*”. It has been argued that the main difference in the computational system of humans, compared to other species, is that capacity for recursion (Hauser et al., 2002). According to this view, the capacity to make recursively nested construction is what allows humans to have such a rich expressiveness in their communications. Indeed, either as speech, sign, text, or Braille, human languages stand out in the animal kingdom with their unbounded combinatory potential.

These are my main motivations to study the neural bases of language, and compositionality most of all. By neural bases, I mean the characterization of the representations and computations supporting language processing and their evolution over time, not simply their localization. Indeed, I believe that, in cognitive neuroscience, localization in space has been extensively studied, and occasionally abused (Logothetis, 2008; G. Miller, 2008), but localization in time can lead to more informative results about the constraints a computational model should have to achieve language processing abilities similar to humans.

The study of the neural underpinnings of language has a rich history. The remainder of this introduction will give a non-exhaustive description of it in a roughly chronological fashion, starting in the 19th century, with the first connections of focal brain lesions to specific linguistic deficits. Then, following with the advent of neurolinguistics in the late 20th century thanks to advancements in neuroimaging techniques, culminating with the description of the neo classic dual-route model of language processing. In parallel, developments in linguistics and natural language processing allowed the formulation of more precise hypotheses about the sub-processes under study. Next, the description of recent studies in two-word compositions, thematic roles, node counts, entrainment to syntactic structure, and finally, semantic mapping using activations from

language models. Ultimately, wrapping-up the introduction is a description of the specific questions that my thesis tried to answer.

B. Early days of language neuroscience: lesion studies

Arguably, the first attempt to localize linguistic processing in the human brain was done by Jean-Baptiste Bouillaud. He found that patients with trouble in understanding or articulating language, a condition known as **aphasia**, often had lesions in the frontal part of the brain (Bouillaud, 1825).

Decades later, Paul Broca achieved a more precise location by systematically comparing the lesions of a cohort of patients with similar symptoms: a deficit in articulating speech but not in understanding it. Broca identified a small section of the left frontal lobe, the inferior frontal gyrus (Broca, 1861, 1865)[1], to be the common factor in their lesions.

This foundational result inspired Carl Wernicke to study linguistic disorders. In turn he discovered that some aphasic patients were actually capable of fluently speaking complex sentences, but devoid of meaning, while having a strong deficit in understanding speech or written text. This condition, quite different from the one Broca described, was associated with a lesion in the temporal lobe, the posterior superior temporal gyrus. In the same publication, Wernicke proposed a model of how these regions interact with one-another (Wernicke, 1874). This is likely the first time a model of language processing in the brain is proposed (Figure 0.1, left). These diseases were later dubbed **Broca's aphasia** (or motor aphasia), and **Wernicke's aphasia** (or fluent aphasia). Broca and Wernicke thus paved the way for others to identify brain areas involved in specific deficits.

Ludwig Lichtheim later described symptoms associated with a lesion in the arcuate fasciculus, a bundle of connections between the inferior frontal gyrus and the posterior

[1] More precisely the posterior two thirds, i.e. the pars triangularis and pars opercularis of the inferior frontal gyrus. Also, it's in the left hemisphere in the majority of humans, but not all.

superior temporal gyrus (i.e. the regions identified by Broca and Wernicke). Patients with such lesions have trouble repeating speech but have intact comprehension. When speaking, they make many mistakes such as substituting or transposing sounds. Lichtheim integrated previous findings with his own by proposing an extension of Wernicke's model of language processing in the brain (Lichtheim, 1885). In this model, the regions identified by Broca and Wernicke are connected to a "concept center", where semantic processing happens (although no anatomical localization, nor mechanism for composition, is proposed). It will dominate the field for more than a century, to the point that it is called the Classic Model of language neurobiology, or the **Wernicke-Lichtheim model** (R. E. Graves, 1997; Nasios et al., 2019). In this model, speech information in the auditory cortex is routed to the posterior superior temporal gyrus, where the meaning of the words is accessed. In the case of reading, information is sent from the visual cortex to the angular gyrus and then to the posterior superior temporal gyrus. When speaking, the meaning of the words is sent from the posterior superior temporal gyrus via the arcuate fasciculus to the inferior frontal gyrus, where morphemes are formed and transferred to the motor cortex. Norman Geschwind is credited with a revival of this model in the 1960s and 1970s (Geschwind, 1965, 1970).

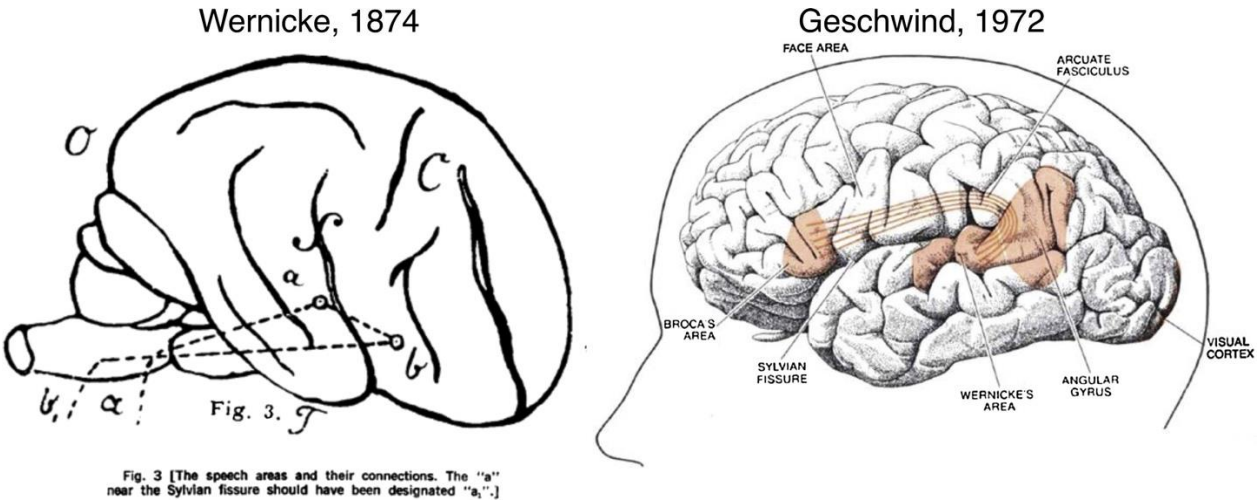


Figure 0-1: Historical models from Wernicke (1874) and Geschwind (1972)

Interestingly, it is very rare to see patients with deficits in elementary compositional processes, such as the ability to form two-word phrases. It seems to

happen only when the whole linguistic ability is impaired, a condition known as global aphasia. This suggests that basic composition is implemented in a robustly distributed, or redundant fashion.

C. The dual-route model of language processing

Much more recently, starting in the 1970s, this influential model has been put into questions by new findings. For example, it was found that patients with lesions in the left superior temporal gyrus (STG) had no deficit in speech comprehension, as would be predicted by the Wernicke-Lichtheim model, but a deficit in speech production (Damasio & Damasio, 1980). This does not completely preclude a role of STG in speech perception, but points to a more complex picture, with more regions involved in the process. Many such findings put the classical model into question and their exhaustive summary would take too much time and space.

More generally, the model was criticized for being too simplistic. Specifically, (i) new aphasic syndromes have been identified that lack an explanation in the classical model, (ii) the terms used to define it do not reflect the advancements in linguistics and thus prevents the fields to interact, and (iii) it is anatomically underspecified, in that the regions historically defined have been found to be heterogeneous (composed of subregions that are anatomically and functionally distinct) and ill-defined (different researchers use the same name to designate different areas) (Poeppel & Hickok, 2004; Tremblay & Dick, 2016).

This called for an update. The emerging dominant model, the new classical model, is the **"dual stream"** model from Poeppel and Hickok (Poeppel & Hickok, 2004; Hickok & Poeppel, 2007). This model proposes that speech processing takes place in a **"dorsal"** and a **"ventral" pathway**. The ventral stream, largely bilaterally organized, encompasses regions from the temporal pole to the basal occipitotemporal cortex. Its role is to map incoming sounds to meaning. On the other hand, the dorsal stream is strongly left hemisphere dominant, from the posterior superior temporal to the inferior frontal cortices. The function of the dorsal route is to map the words (that one plans to say) to

their constituting phonemes, and in turn to the corresponding motor articulations (Saur et al., 2008).

Importantly, this model helped moving from a purely localizationist framework, pervasive in cognitive neuroscience, to a computationalist perspective, in which the end-goal is not to localize a brain function but to achieve a mechanistic understanding of how it is achieved. As Poeppel and Hickock put it: "*to have theoretically precise, computationally explicit, biologically grounded explanatory models of the human brain's ability to comprehend and produce speech and language*" (Hickok & Poeppel, 2007).

This model has been validated in meta-analyses (Mirman et al., 2015). Subsequent updates and extension of this model are a reference nowadays (Friederici, 2012; Poeppel et al., 2012; Matchin & Hickok, 2020).

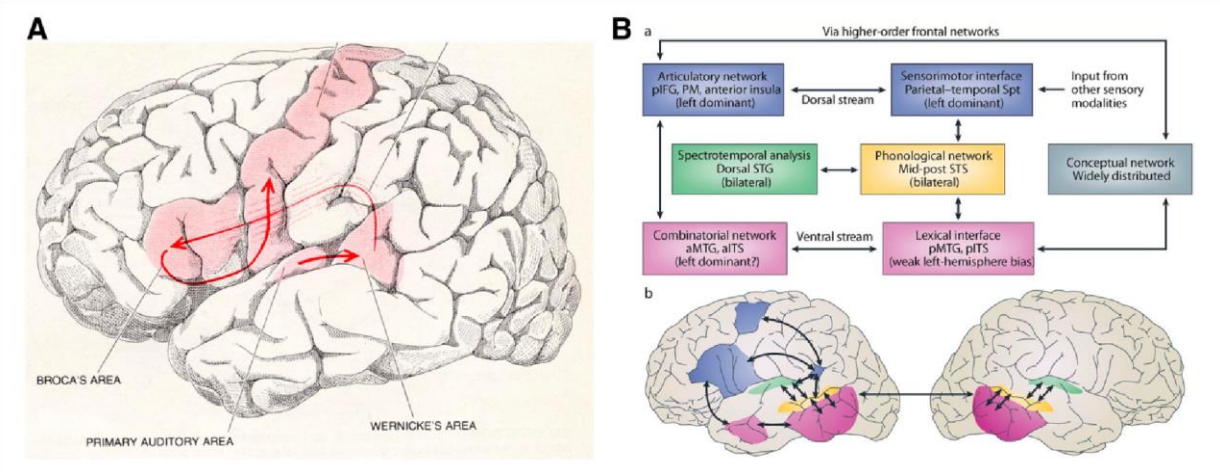


Figure 0-2: Side-by-side comparison of the Geschwind model to the dual-route model (from Poeppel et al 2012)

A: Historical Geschwind model
B: Hickock and Poeppel's dual stream model

Regarding our overarching question: "how are words combined to create sentential meaning?", in this model we find the "combinatorial network", part of the ventral stream. Most importantly, the anterior temporal lobe (ATL) has been associated with semantic composition, by studies showing its higher activations in response to sentences, compared to word lists (Friederici, Meyer, et al., 2000; Humphries et al., 2001;

Vandenberghe et al., 2002), then more recently its involvement in two-word composition, as described in a later part of the introduction.

These theoretical advancements were allowed thanks to two kinds of increase of "resolution" (Poeppel et al., 2012). The first is the advancement of neuroimaging that granted a better **temporal and spatial resolution** than lesion-based deficit mapping. The second is the improvement in "**conceptual resolution**", namely a tighter bridge with linguistics and computational modeling. These have given language neuroscience a stronger theoretical foundation, and made some questions computationally explicit, allowing more fine-grained experimental testing. Even more recently, machine learning and natural language processing have given a third opportunity for progress: an increase in "**computational resolution**". For the first time, we have access to *in silico* models that actually process natural language and perform related tasks at a level comparable to humans. This opens many new opportunities for testing hypotheses about language processing. The next section details such advancements in linguistics and natural language processing that led to innovative experiments in neuroscience.

D. Advances in linguistics

It is now clear that the coarse distinction between comprehension and production in the classical model is very much underspecified. Thankfully, linguists have been making huge progress in formally describing natural language. For example, the theoretical categorization of linguistic subdomains, such as "lexical access", or "thematic roles" allows for a better granularity in experimental designs.

Of interest, the generative grammar proposed by Chomsky (Chomsky, 1957) aims at finding a small set of rules that explains the immense set of well-formed expressions for a given language. This was a drastic change, in sharp contrast to both the previously dominating paradigm in linguistics: **structuralism** (Harris, 1951), and, more generally, the main paradigm in cognitive science at the time: **behaviorism** (Quine, 1960; Skinner, 1957). It put in the spotlight the ability to generate an unlimited supply of hierarchically structured expressions, rather than the inventory of a fixed body of knowledge, as in structural linguistics. Thus, it opened-up the possibility to look for the internal

mechanisms encoded in the brain that underlie such generative processes. This led to the creation of "**the biolinguistics program**", the first branch of linguistics that explicitly looked for the biological underpinnings of linguistic ability, both in terms of biology and evolution (Lenneberg, 1967).

This opposition to behaviorism was made obvious by the famous poverty of stimulus argument (Chomsky, 1965). Briefly, Chomsky argued that the quantity of speech that infants perceive during their early years isn't enough to learn language from scratch, thus the brain has to have **innate components** tailored to learn language. In other words, the brain must be pre-wired for language. This argument has been controversial since its introduction (Piattelli-Palmarini, 1980; Legate & Yang, 2002) and has been debated ever since (Berwick et al., 2011, 2013; Lasnik & Lidz, 2016). It has seen a revival in recent experimental designs that try to test it in humans (Wilson, 2006), as well as in artificial neural networks models of language processing (Warstadt & Bowman, 2022).

Also of note, the idea of a critical period of development for language was a strong argument against pure behaviorism because it argued that language acquisition is biologically constrained (Penfield, 1959; Lenneberg, 1967).

By way of explanation, the generative program put **syntax** first, as opposed to phonology and morphology, center to structuralism. It examines how people parse different constructions to extract their meaning, and serialize a thought into a linear sequence to convey its meaning to other people. At its center is the **grammar**, i.e., the set of rules that stipulates authorized transformations to generate a valid sequence (Greibach, 1978; Savitch, 1987; R. Jackendoff, 2002). Thus, generative systems depict the generation of an infinite range of expressions from a set of elementary elements, using a finite set of rules. The composition operation is therefore assigned a paramount role in the generative framework.

From this paradigm shift stemmed many conceptual innovations, such as the description of **thematic roles** (R. S. Jackendoff, 1972), that interpret the links between verbs and their arguments. These relationships can be viewed as semantically informed

syntactic constraints on the structure of sentences, and include among others the roles of *Agent*, *Experiencer*, *Instrument* and *Goal*.

This tradition culminated in the 1990s with the proposal of the **"minimalist program"** (Chomsky, 1993, 2013). Center to it is the idea of **"merge"**, a general-purpose, binary, and recursive operation. Basically, it takes two syntactic objects, that can be lexical elements or the results of previous merge operations and combine them in a new object that inherits its label from one of the original objects, called the head of the resulting phrase. It is proposed that the structure of syntactic trees can be generated from the recursive application of this operation. In this sense, recursion means repeatedly applying a unique function to its own output. Complicated grammars can thus be replaced by a single universal operation, with similar tree-structure building properties (Dehaene et al., 2015). This parsimonious theory has triggered interest from neuroscientists (Boeckx, 2013; Fukui, 2017).

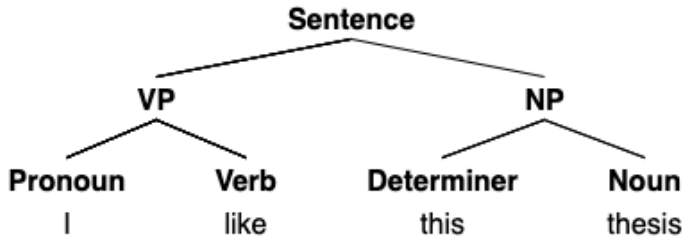


Figure 0-3: Syntactic trees can be built with the recursive merge

To build the syntactic tree of this sentence, three merge operations would be necessary. First to combine the pronoun "I" to the verb "like" in a Verb Phrase (VP). Second, to make a Noun Phrase (NP) out of the determiner "this" and the noun "thesis". Finally, the VP and NP have to be merged to produce the whole sentence "I like this thesis".

These conceptual shifts were necessary conditions for researchers to be able to formulate hypotheses about the processing of language in the brain. The compositional operation has now been characterized theoretically, paving the way for neuroscientists to study its neural underpinnings. The search for the neural basis of composition can now start. The next section will detail this neuroscientific adventure.

E. Studies on linguistic composition in the brain

1. Technical developments

Considerable progress has been made in the neurobiology of language in the recent past. On one hand, these advances are due to more precise hypotheses formulation, thanks to development in linguistics described in the previous section. In tandem, major technical improvements in the neuroimaging methods allowed better and better spatial and temporal resolutions, which we summarize here. Jointly, these breakthroughs empowered us - and still do nowadays - to test evermore precise hypotheses about language processing in the brain.

This enterprise started in the mid 20th century with two main developments. The first is the use of electrical stimulation of the exposed cerebral cortex during surgical operations in awake patients, led by the pioneering work of Wilder Penfield (Penfield & Rasmussen, 1950; Penfield, 1959; Manuel, 1979; Isitan et al., 2020). This paved the way for the extensive mapping of language areas using similar techniques (Ojemann et al., 1989). The second has been dubbed the Wada Test and consists of an injection of barbiturates in the carotid artery on one side, resulting in a transient hemispheric inactivation, allowing to test for hemispheric language dominance (Wada, 1949; Branch et al., 1964; Geschwind, 1970).

But major changes were yet to come. The development in the 1970s of the computed tomography and magnetic resonance imaging (MRI) scans allowed to map lesions in living patients (Hounsfield, 1980). Building on the later, Seiji Ogawa created the functional MRI (Ogawa et al., 1990), which triggered in the 1990s a revolution in cognitive neuroscience (Moonen et al., 1990; M. S. Cohen & Bookheimer, 1994; DeYoe et al., 1994). Based on the hemodynamic response, it allowed for the first time to non-invasively quantify brain activity with high spatial resolution. This permitted to study language processing in healthy subjects, by comparing brain activations (measured as changes of blood flow) to linguistic tasks and comparing them to closely related control tasks, such as listening to a known versus unknown language, or sentence reading versus word lists reading (Mazoyer et al., 1993; Binder, 1997; Binder et al., 1997; C. J. Price, 2000). Of

interest, it was found that hearing meaningful sentences activates the left middle temporal gyrus, the left and right temporal poles, and a superior prefrontal area in the left frontal lobe (in addition to regions devoted to single word lexical access) (Mazoyer et al., 1993). Also noteworthy, comparing a sentential description to a closely matched environmental sound description, it was discovered that the sentence elicited more activation in the anterior temporal cortex (Humphries et al., 2001), suggesting that it is linked in sentence processing, somewhat devoid of the semantic content. A followup study using pseudowords showed that, indeed, the left anterior temporal lobe (ATL) activity is driven more by syntactic cues than semantic ones, whereas the angular gyrus, while being modulated by syntactic structures, is driven more by semantic content (Humphries et al., 2006). Notable as well, fMRI studies started charting a semantic map in the brain, characterizing a set of temporal and frontal areas (Binder et al., 2009; Binder & Desai, 2011).

In parallel, multichannel electroencephalography (EEG) (Berger, 1929) was being democratized and magnetoencephalography (MEG) was invented (D. Cohen, 1968; Sato & Smith, 1985). These methods have an excellent temporal resolution, thus allowing researchers to precisely chart the time course of neural activations during language processing. It also allowed the study of the role of neural oscillations in language processing, thus linking the field with systems neuroscience and electrophysiological research in animal models (Buzsáki & Draguhn, 2004). For example, it was discovered that lexical access was coupled with an increase in the theta rhythms (4-7 hertz), whereas composition operations induce increase in the beta (12-30 hertz) and gamma (over 30 hertz) frequency bands (Bastiaansen & Hagoort, 2006).

Many more cutting-edge technologies have been put forward in recent years (Seo et al., 2016; Bihan & Schild, 2017; Musk & others, 2019; Steinmetz et al., 2021). There is no doubt that the ever increasing quality of brain recordings both raises new challenges and offers unrivaled opportunities to test brain function (Urai et al., 2022), as I will explore in the discussion.

In the following sections I review recent literature related to semantic processing, with a focus on the temporal dynamics, setting the stage for the presentation of my own work in the coming chapters.

2. A window into sentence processing: the N400 and P600 components

The electrophysiological study of language processing begins with the discovery of the now famous N400, a component of evoked related potentials (ERPs) consisting of an increased negativity in response to semantic violation such as in "I like my coffee with cream and *socks*" (Kutas & Hillyard, 1980). Since this foundational study, an enormous body of work focused on this component, finding it present in response to unexpected stimuli not only in language processing, but also in object, face, action, gesture processing and mathematical cognition (Kutas & Federmeier, 2011). Although its interpretation has been steadily debated since its discovery, there is some consensus on the fact that the amplitude of the N400 is modulated by the amount of surprise: the less predictable the stimulus is, the bigger the negativity (Lau et al., 2008).

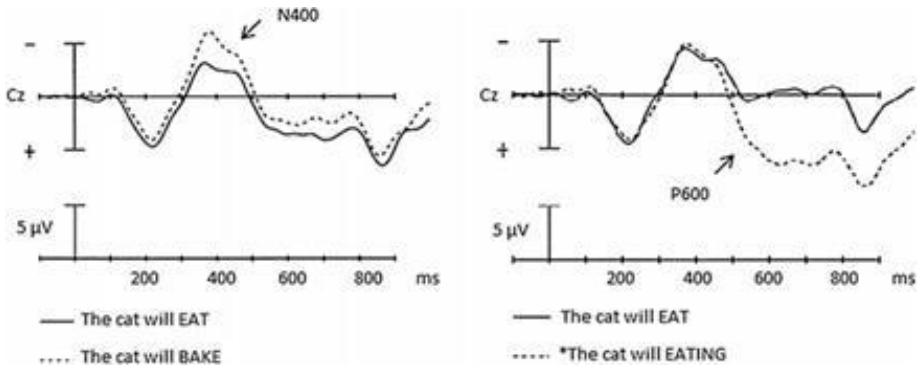


Figure 0-4: N400 and P600 components along with examples sentences that can elicit them

Voltage measured at the Cz electrode (situated on the midline, at the center of the scalp), in response to normal (full line) and erroneous (dashed line) sentences. As a convention, negative voltage is plotted upward. Taken from Osterhout, 1991.

A few years later, researchers discovered that the introduction of syntactic ambiguities elicited a later positivity, dubbed P600 (Osterhout, 1991), and thought to be linked to syntactic integration (Guillem et al., 1995; Kaan et al., 2000). This simplistic view of N400-semantic and P600-syntactic has since been put into question (Frenzel et al., 2011), but no consensus seems to arise. Examples of recent models of these phenomena include the Retrieval-Integration model, in which the N400 component reflects the access of word meaning in memory, and the P600 component marks the integration of the word into the developing phrasal representation (Brouwer et al., 2017; Delogu et al., 2019). In another proposal the N400 amplitude indicates the adjustment brought about by an incoming word in a probabilistic representation of meaning (Rabovsky et al., 2018).

3. Neural signatures of two-word compositions

Of special interest to us, a series of studies focused on the time-resolved analysis of two-word composition in MEG (reviewed in (Pylkkänen, 2019, 2020a)). They showed that the composition of a noun and an adjective into a noun phrase (e.g., “red boat”) elicit a transient increase in activation in the left anterior temporal lobe (ATL) around 200–250 ms, and in the ventromedial prefrontal cortex around 400 ms after the onset of the second word (Bemis & Pylkkänen, 2011). They used word list controls where no composition could have happened (e.g., “cup boat”). Besides reading, this effect seems to also be present in hearing (Bemis & Pylkkänen, 2013b) and in production of similar two-word English phrases (Pylkkänen et al., 2014). It is also present in Arabic (Westerlund et al., 2015) and in American Sign Language (Blanco-Elorrieta et al., 2018). Similar effects were found for verb phrases (Kim & Pylkkänen, 2019) and noun-noun compounds (Brooks & Cid de Garcia, 2015; Flick et al., 2018). This experiment was later replicated in EEG and found a greater centroparietal negativity between 180-400ms for words in a context where composition could happen (Neufeld et al., 2016). This timing and localization, akin to that of the classical N400 effect, suggests that they could have similar neural origins.

In addition, it was found that the activation in the left ATL was linked to the specificity of the concept that is being processed. Here, specificity or generality are functions of the size of the set of elements that belong to the category; “*cat*” is more

specific than "*mammals*", because cats are a subset of all mammals. The activity in left ATL was found to be higher for a more general head (noun) and a more specific modifier (adjective). In other words, the bigger the reduction in the space of possible concepts, the bigger the activation in left ATL (L. Zhang & Pylkkänen, 2015). For example, "*tomato dish*" elicited higher activity than "*vegetable dish*", because "*tomato*" is a more specific modifier than "*vegetable*" (see figure 0.5), and "*red boat*" elicited a higher activity than "*red canoe*", because "*boat*" is a more general head than "*canoe*". Finally, it was found that the modifier doesn't need to carry any feature in itself: the effect is still present with the adjective "same" that has the property of referring to other features in a specific context. In this case, if "same" referred to both the color and size (of a preceding picture), it elicited a higher activation than when it referred to a single feature (L. Zhang & Pylkkänen, 2018).

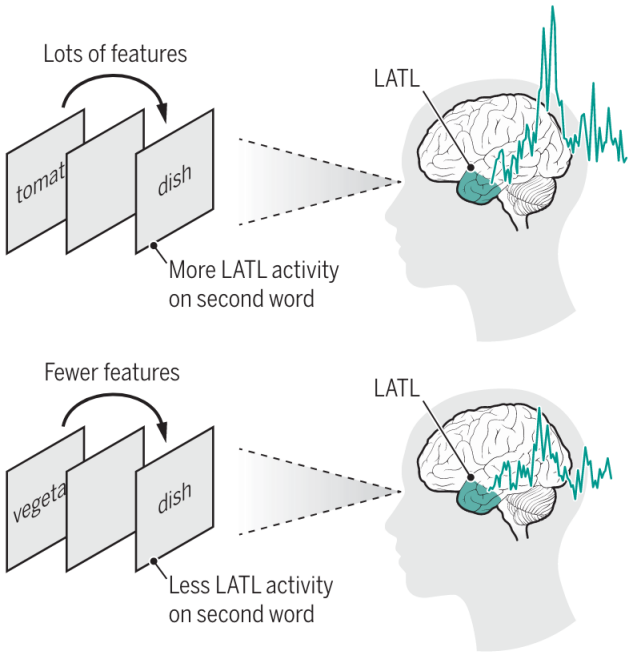


Figure 0-5: Left ATL sensitivity to conceptual specificity

All this suggests a modality independent function in conceptual composition of the left ATL, as well as some later (~ 400 ms) implication of the ventromedial prefrontal cortex, and the angular gyrus. It is interesting to note that this late frontal effect in comprehension is reversed during production, where the prefrontal cortex is now activated first (150-200 ms for a picture naming task), suggesting that we are witnessing

the planning phase of sentence production (Pylkkänen et al., 2014). This suggests that the ventromedial prefrontal cortex is placed higher in the language hierarchy: it may contain compositional representations, either after their processing (in comprehension), or before they are linearized into lexical items for production.

In another line of work using fMRI, activations in the left ATL were found to be compositional: the activity evoked by a concept in a voxel could be linearly predicted from the activity of its semantic dimensions (e.g., the representation "boy" can be predicted from the representations of "young" and "man") (Baron & Osherson, 2011). These neuroimaging findings, as well as neuropsychological and transcranial magnetic stimulation (TMS) results (Jefferies, 2013), make left ATL the most likely locus of a "conceptual" form of composition.

Interestingly, another study found that, in a similar two-word composition paradigm, the adjective stays represented explicitly (in whole brain MEG signals) when it is waiting to be composed with a following noun (Fyshe et al., 2019). This phenomenon is explored in our second study, but in French (as opposed to English) and with phrases of varying length.

Complementing these numerous MEG studies, better spatially resolved fMRI investigations of two-word composition also strongly implicated the inferior frontal gyrus (IFG), especially the pars opercularis (Brodmann Area 44) (Zaccarella & Friederici, 2015; Schell et al., 2017).

4. Studies on increase of sentence and node counts

This wealth of research on two-word compositions has certainly been fruitful (Pylkkänen, 2020a), but concerns can be raised about the ecological validity of such findings (Varoquaux & Poldrack, 2019; Willems et al., 2020). Are the signatures identified in this simplified context representative of how "merging" happens during natural language use? In this section we go one step further and consider whole sentences. In some studies, syntactic features are also considered as predictors of brain activity, since

linguistic theory predicts that different computations must be done depending on the syntactic structure of a sentence.

An MEG study showed an increased activation for sentences, compared to word lists, in ATL but also in posterior temporal, inferior frontal, and ventral medial areas (J. Brennan & Pylkkänen, 2012). What's more, this activity correlated with syntactic features, as quantified by the number of steps of a left-corner parser (J. R. Brennan & Pylkkänen, 2017). Numerous fMRI studies have also found increased activation in the left ATL and the left IFG when comparing sentences to word lists (Mazoyer et al., 1993; Stowe et al., 1998; Friederici, Meyer, et al., 2000). Interestingly, in some cases this increase was also present when the stimuli used were not real, but pseudowords (Humphries et al., 2006).

But to what exactly is this increase due? It should not be linked to working memory load, because it is constant between sentences and word lists. We turn to linguistic theory for a new hypothesis: it is thought that the merging operation can take place only when a constituent can be meaningfully closed. In this proposal, words waiting to be composed are stored in a buffer, then merged at the end of the constituent. The merged representation is believed to be compressed, and thus characterized by a decreased activation. In the end, this proposal predicts a linear increase in activation with the number of elements in the sequence that can be combined together (e.g., words in a phrase, but not in a list), followed by a decrease.

In a foundational study, Pallier and colleagues tested this hypothesis and found an increase of activation evoked by phrases of increasing length (compared to word lists of similar lengths) in the Superior Temporal Sulcus and the pars triangularis and orbitalis of the left IFG, as well as in the temporal pole, and temporo-parietal junction (Pallier et al., 2011). Interestingly, the increase in activation with each additional word was logarithmic, not linear (except in the temporal pole). Crucially, in the posterior temporal sulcus and the IFG, the effect was still present when normal words were replaced by meaningless pseudowords, suggesting a role in the purely syntactic processes: the computation of the constituent structure of a phrase can happen even in the presence of meaningless content. On the other hand, the anterior superior temporal sulcus, temporal pole and

temporo-parietal junction were only more active during normal sentences, suggesting a tight role in (creating or maintaining) semantic representations.

This study was followed-up by Nelson and colleagues, who used more varied syntactic structures (not only simple left branching ones) to confirm that the increase of activity following each word in a constituent would decrease when the constituent could be meaningfully closed (i.e., all words can be merged into a more compact representation), then increase again with the next constituent's incoming words (Nelson et al., 2017). In other words, if humans indeed "parse" sentences, as defined by linguists, there should be some neural marker of the storage of the elements of a constituent waiting to be merged together (also called number of open nodes in the syntactic tree, see figure 0.6 A), and a transient marker of the merge operation after the end of the constituent. The study used intracranial recording, which possesses excellent spatial and temporal resolution, but where the placement of electrodes depends on clinical, not scientific motivations. As a measure of neural activity, they computed high gamma (70 - 150 hertz) power, which is broadly accepted as reflecting the firing rate of neurons close to the recording site (K. J. Miller et al., 2009; Ray & Maunsell, 2011). They found that in the superior temporal and inferior frontal cortices, high-gamma power increased by a fixed amount with each successive word (Figure 0.6 B) and, as soon as a phrase could be meaningfully closed, an additional burst of activity happened, most notably in the pars triangularis of the IFG (Figure 0.6 C). Quickly after, the high-gamma power dropped, echoing a "compression" of the merged elements into a single unified phrase.

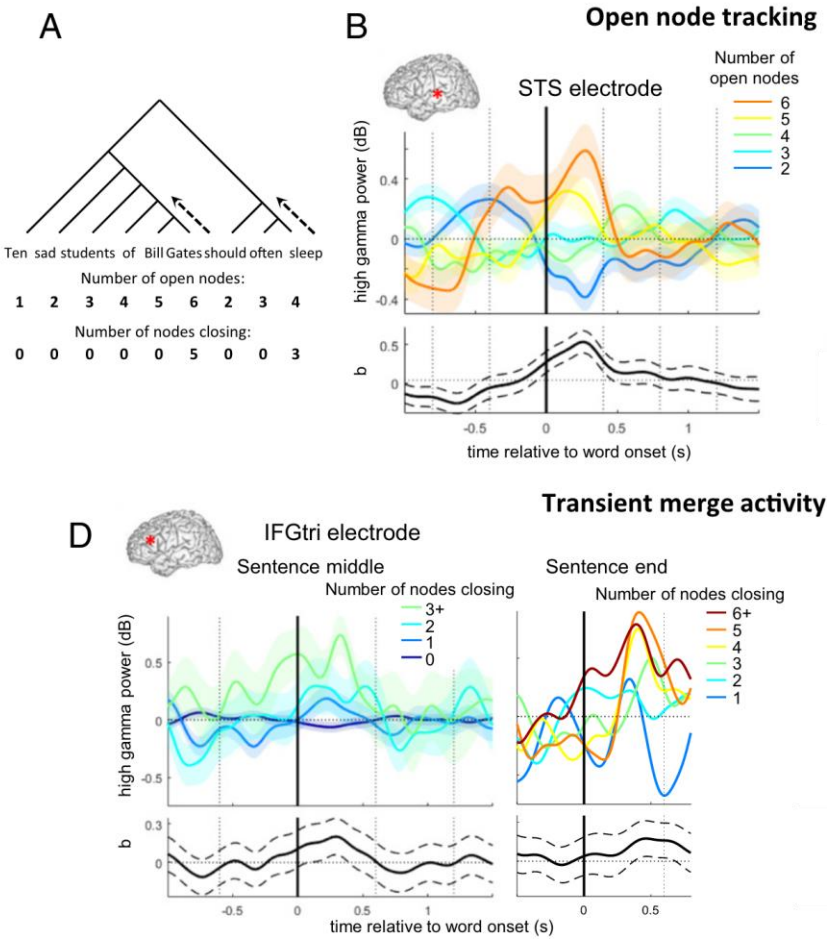


Figure 0-6: Open node tracking and transient merge activity in single intracranial electrodes (from Nelson et al. 2017)

These findings are also corroborated by another intracranial study that found a similar build-up effect over the course of the sentence (Fedorenko et al., 2016). Interestingly, the effect was also somewhat present, although weaker, for word lists, and weaker still for sentences made of pseudowords (*aka* Jabberwocky). This could suggest that semantic and syntactic dimensions are recruited independently: in word lists only the neural assemblies carrying the meaning of individual words would be recruited, while in Jabberwocky only the assemblies that bear syntactic information. In normal language, both these assemblies would be recruited, as well as compositional constructs resulting from the interaction of the two. Additional support for the separation of structural (syntactic) and meaningful (semantic) in the brain sprouted from high-resolution fMRI studies: it was found that the pars opercularis of the IFG (Brodmann Area 44) responded

to syntactic, but not lexical information, while the pars triangularis (Brodmann Area 45) is active only if semantic information is present (Goucha & Friederici, 2015).

Notably, Brennan and colleagues found that the activation in the left ATL, but not lIFG, correlated with the number of open nodes of a syntactic parsing during natural story listening (J. Brennan et al., 2012). This setup, although less controlled than the preceding experiments, confirms the involvement of the left ATL in semantic composition in a realistic setting. In the next sections we develop two sets of studies on thematic roles and entrainment to syntactic structure that provide a complementary view on meaning composition.

5. Thematic roles studies

Another way to tackle the problem of compositionality is through the prism of thematic roles, i.e., "who did what to whom" in a sentence (see the section on Advancement in linguistics for a more detailed account). Indeed, single words can be (in different contexts) both subject and object of an action, e.g., a person can both give and receive a kiss (also, possibly, both). This raises the question of how such important characteristics are encoded in the brain. What are the differences between the representations of "*Marie kisses John*" and "*John kisses Marie*"? In a foundational study, Frankland and Greene (Frankland & Greene, 2015) tackled this question with multivariate decoding in fMRI and identified a region in left mid-superior temporal cortex where thematic roles could somewhat be identified in different subregions (Figure 0.7). In other words, neighboring subregions weakly but consistently encoded the values answering the questions "Who did the action?" and "To whom was it done?". These results are corroborated by lesions studies, in which patients with lesions to the left mid-superior temporal cortex have a specific deficit in identifying thematic roles (Wu et al., 2007).

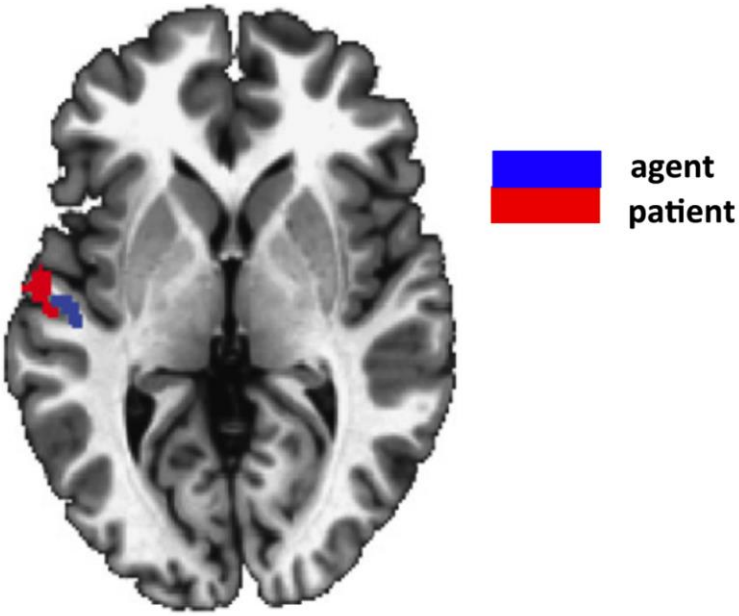


Figure 0-7: Regions identified as encoding the identity of the agent and patient of a sentence (from Frankland & Greene, 2015)

In a follow-up study (Frankland & Greene, 2020b), they examined the effect of role specificity/generality (in the sense that "agent" is the most general, because it encompasses all possible agents and, e.g., "chaser", being a specific kind of agent, is more specific). Using encoding models, they identified a region in the anterior-medial prefrontal cortex that encoded noun-verb conjunctions such woman-as-chaser, with a distinct representation compared to woman-as-chasee. In contrast, the left-mid superior temporal cortex encoded only general roles (agent, patient). Finally, the hippocampus was found to encode events that shared narrow roles as more dissimilar, compatible with a role in pattern separation (Yassa & Stark, 2011). This gives a broad picture of the encoding of thematic roles, with temporal and frontal cortices having complementary roles (Frankland & Greene, 2020a).

These findings back up the studies on simple composition described above. They confirm an important role of the temporal and frontal lobes in conceptual composition. They also suggest a resemblance to a classical computer, where data registers store the temporary value of a variable, allowing to flexibly generate combinations of variables on the fly. However, this raises a scaling issue: there can't be as many subregions as there are

possible roles. It is sensible that roles as important as “subject” and “object” of the action have a dedicated region to store their value, but further specialization raises yet unanswered questions.

6. Entrainment to syntactic structure

A set of studies tried to go beyond two-word compositions using a paradigm based on neural entrainment. The rationale behind entrainment is that by repeating stimuli at a given frequency, the neural activity will lock onto it and generate rhythms at the same frequency, detectable by time-frequency analysis. Critically, if the brain is also sensitive to subharmonics present not in the sensory input, but in a higher-level structure present in the stimuli, this should be reflected in the neural signals as well. For example, in music, entrainment to the overall meter was found in cases where only a few beats per bar were present (Nozaradan et al., 2012). This paradigm have been used extensively in the study of speech processing (Ding & Simon, 2014; Obleser & Kayser, 2019), for example many auditory regions have been shown to be entrained to the envelope of the acoustic signal, and this entrainment was shown to modulate speech intelligibility (Vanthornhout et al., 2018).

In a groundbreaking study, Ding and colleagues tested the neural entrainment to syntactic structure (Ding et al., 2016). They used simple 4-word sentences composed of a two-word noun phrase and a two-word verb phrase (Figure 0.8 a). Words presented at a 4 hertz frequency elicited entrainment at 4 hertz, but also at 2 hertz (the phrasal rate) and 1 hertz (the sentence rate, Figure 0.8 c), although these subharmonic were not present in the stimuli power spectra (Figure 0.8 b). This result provided a long sought for neural signature of syntactic structure.

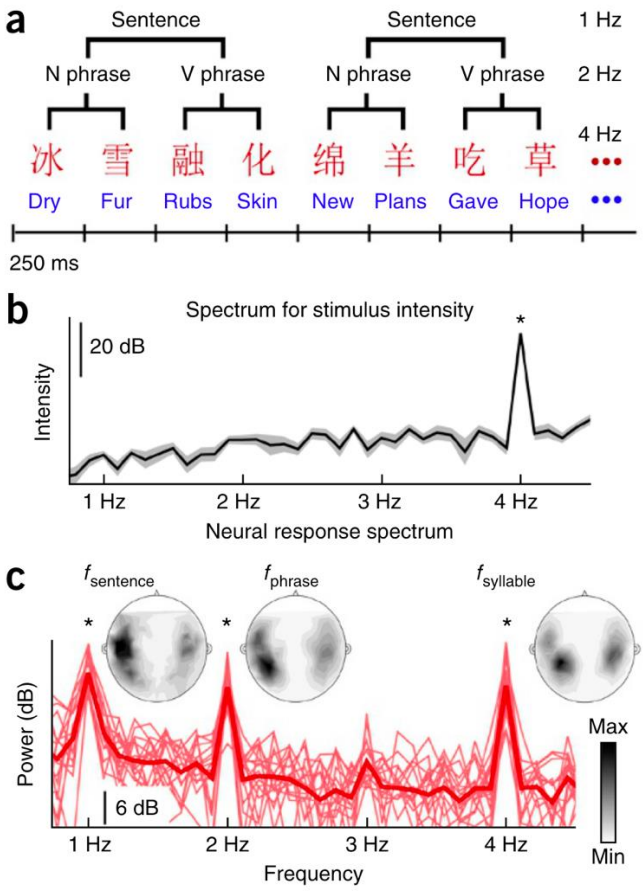


Figure 0-8: Neural tracking of hierarchical linguistic structures

a: stimuli used and their presentation rate.
b: power spectrum of the stimuli
c: power spectrum of the neural data

Doubts were raised by subsequent studies that claimed that the verb rate (equal to the sentence rate) could be the source of the entrainment (Frank & Yang, 2018; Tavano et al., 2020), thus suggesting that the finding is lexical, not syntactic in nature. However, these doubts seem to be vanishing thanks to a follow-up study that showed that word lists, containing the same lexical elements but without sentential structure, did not elicit entrainment (Lo et al., 2022).

All-in-all, previous studies provide evidence that parts of the medial prefrontal cortex, inferior frontal gyrus, angular gyrus, left mid-anterior superior temporal, and

superior temporal sulcus cortex contribute to the construction of conceptual representations out of simpler parts. This concludes the presentation of the neuroimaging literature that looks for neural correlates of high-level linguistic properties. Next, we introduce a new kind of model of language processing based on artificial modeling.

7. Advances in Natural Language Processing

Much more recently than the generative endeavor in linguistics, progress in another field also allowed testing brand new hypotheses in the cognitive neuroscience of language: that is computational linguistics and natural language processing (NLP). The very recent explosion in popularity of **artificial neural networks (ANNs)** for NLP stems from the longstanding distributional hypothesis, which can be summarized as follows: "words which are similar in meaning occur in similar contexts" (Rubenstein & Goodenough, 1965). Namely, words that are found in similar environments must share some properties, whereas words that are found in different situations ought to have nothing in common. For example, color adjectives (blue, red, yellow, …) are very similar under the distributional hypothesis because they can be used interchangeably in many cases. On the other hand, adjectives and adverbs are very dissimilar because, as belonging to different syntactic categories, they happen at different places in a sentence. This intuitive approach, however, does not stipulate how the similarity of the words can be measured. The trick to go from this theory to functional NLP methods was to go to distributed, continuous spaces. Famously, the word2vec algorithm used an ANN to predict a word given its surrounding context, and yields a vectorial representation that substantially captures the word's meaning (Mikolov, Sutskever, et al., 2013; Mikolov, Yih, et al., 2013), called a **word embedding (WE)**. For example, the male/female relationship is well represented in this space using simple vector arithmetic:

$$word2vec(\text{“}King\text{”}) - word2vec(\text{“}Man\text{”}) + word2vec(\text{“}Woman\text{”}) \approx word2vec(\text{“}Queen\text{”})$$

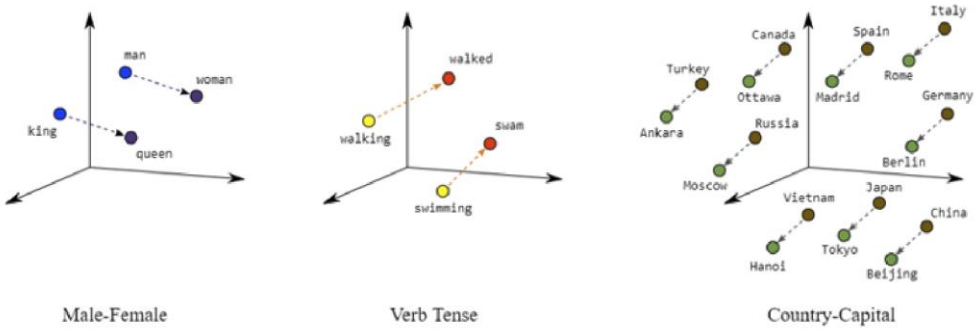


Figure 0-9: Examples of interpretable directions in the embedding space of word2vec

These WE revolutionized NLP, as they capture many important lexical properties and are learned end-to-end from large corpora. They are a genuine embodiment of the distributional hypothesis. Following this foundational work, much progress has been made to improve vectorial word representations. The main gain in performance for many practical NLP tasks (sentiment analysis, document classification, summarization, …) was the appearance of dynamic, or contextual word embeddings (contrasting them with their static ancestors). In **contextual word embeddings (contextual WE)**, the representation of a word also depends on its surrounding context. Thus, a word will have a different embedding for each possible context it can be found in, whereas the classical, static WE assign to each word a unique vector. Consequently, CWE can help disambiguate polysemous words: for example, the word “hot” will get different CWE in the sentences “*It’s very hot today*” and “*He’s so hot*”, reflecting their difference in meaning. Another example can be found in Figure 0.10. Using an analogy serving our overarching question of compositionality, one can liken static WE to lexical semantics, i.e., the meaning of a single word, and contextual WE to compositional semantics, that is the joint meaning of multiple words when they are presented together.

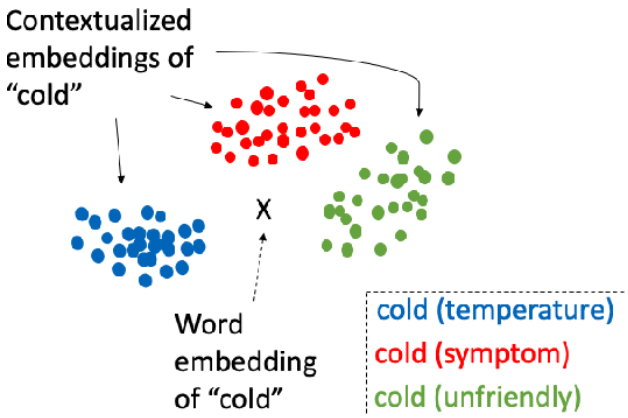


Figure 0-10: Contextual word embeddings can disambiguate the meaning of polysemous words, but static word embeddings cannot

Contextual WE were first described using **recurrent neural networks (RNNs)**, such as simple RNNs (Elman, 1990) and gated variant like Long-Short Term Memory (LSTMs) and Gated Recurrent Units (GRUs) (Cho et al., 2014; Hochreiter & Schmidhuber, 1997). These networks are trained with backpropagation through time to predict the next word in a sentence from the preceding ones, the so-called **language modeling** objective, and showed significant improvement over word2vec-like approaches (Li, 2022).

Then, in 2017, a new architecture based on a self-attention mechanism was proposed: the **Transformer** (Vaswani et al., 2017). Instead of taking the input one word at a time like a human or a RNN, the Transformer takes the sequence all at once, and its multiple attention heads focus on different parts of it. This operation, repeated across multiple layers, extracts powerful linguistic properties. Thus, in theory, even long-range dependencies can be easily captured by a Transformer. Furthermore, the massively parallel architecture makes for easier training and better scaling, allowing bigger and bigger models (Radford et al., 2018, 2019; Brown et al., 2020a; Raffel et al., 2020; S. Smith et al., 2022). Of interest to us, this gain in performance is - on the surface - at the cost of a loss in biological realism: humans process text and speech sequentially, one word at a time, not all at once like a Transformer.

Parallel to the increase in size, a new objective function was proposed to allow bidirectional context to be taken into account. Instead of the classical next-word prediction task, researchers developed a variant where both left and right context are

passed to the model, but some words are masked, and the model has to reconstruct the full sentence. This so-called “masked language modeling” objective allowed further improvements in various tasks (Devlin et al., 2019). Again, the resulting representations, although better for NLP benchmarks, are no longer a reasonable model for how humans process language: the bidirectional contextual embeddings contain information about words in the future.

However, these deep learning-based models are the current best behaving models of language processing. They are able to reach human-level performance on many tasks (Storks et al., 2019; Tripathy et al., 2021) and can even write coherent articles given only a simple prompt (GPT-3, 2020, p. 3) and thus, in this aspect, are unlike any other theories of language processing (e.g., (A. E. Martin, 2020)). Finally, contextual WE from Transformers are currently the best predictors of brain activations to linguistic inputs and are used for hypothesis testing in many recent neurolinguistic experiments, which we will describe in the next section.

8. A semantic map for the brain

We have previously described studies that looked for neural correlates of linguistic constructs: the “merge” operation, constituent structure, parsing operations and thematic role assignment. To conclude this introduction, we present studies that used internal representations of deep neural language models to look for where in the brain is compositional semantics encoded. Indeed, as described in the preceding section, static WE are good vectorial representations of lexical semantics, and contextual WE from neural language models are a rich vectorial representation of compositional semantics. Consequently, they provide a one-of-a-kind opportunity to study meaning composition, and as such researchers have recently started to align them to neuroimaging data.

Specifically, unless otherwise stated, these studies used encoding models in the form of regularized linear regression, to predict brain activations (voxels or electrodes) from different kinds of word embeddings. The performance of the encoding model was quantified with a correlation between predicted and actual brain data, computed on a

held-out set not used for fitting the model's parameters, also called brain score (Schrimpf, Kubilius, Hong, et al., 2020).

The first study of this kind used a small number of concrete nouns and "simple" cooccurrence vectors (based on the co-occurrence of the word with 25 hand-defined verbs). The encoding models successfully generalized to nouns unseen during training (J. Mitchell & Lapata, 2008), suggesting for the first time that continuous vectorial representation could provide good features for brain encoding models.

The first study to use natural (audiobook) stories and end-to-end trained WE with fMRI showed that semantic activations are distributed all over the cortex, with strong grouping and regional specialization (Huth, de Heer, et al., 2016). For example, most activation evoked by "numeric" concepts were grouped together around the parietal cortex, whereas "social-emotional" concepts are found in multiple places, but always clustered together (Figure 0.11). Interestingly, this semantic map was relatively symmetrical across the two cerebral hemispheres. This study thus suggests that semantic networks encompass the whole cortex, contrasting with previous beliefs that language is very localized in the brain. A follow-up study confirmed these results while including spectral and articulatory covariates in the encoding model (Heer et al., 2017). It was also later shown to be invariant to stimulus modality (speech or text) (Deniz et al., 2019), confirming that the brain activity predicted by WE are amodal.

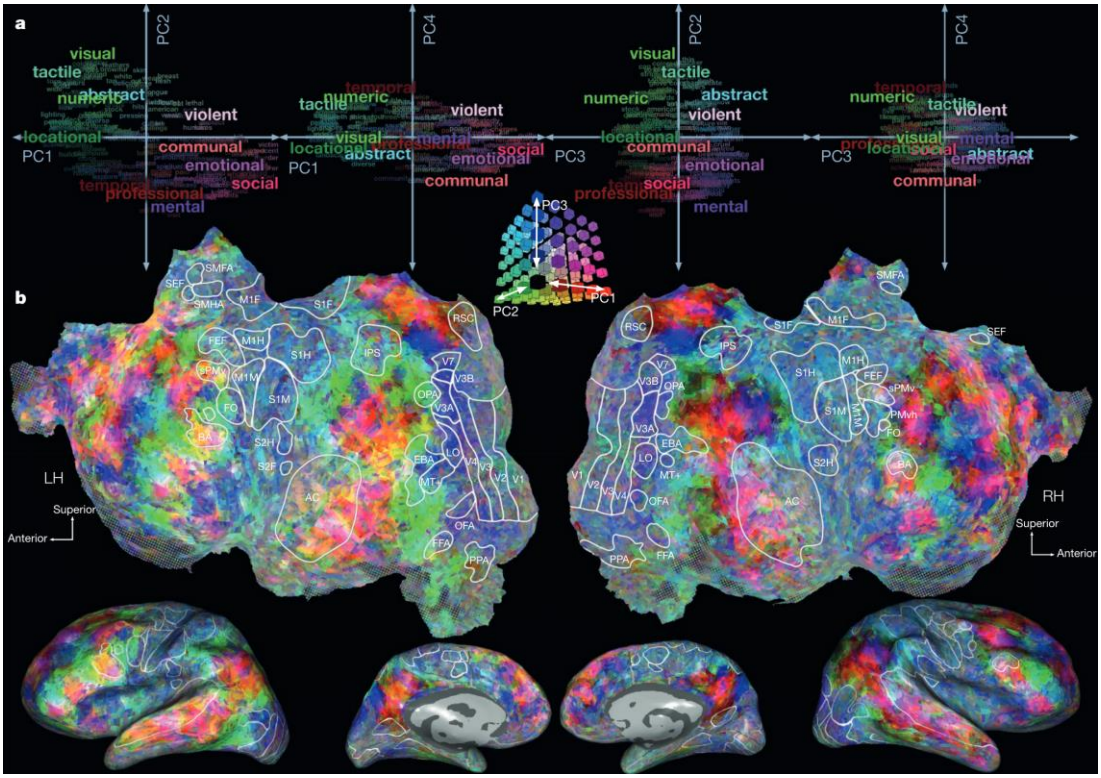


Figure 0-11: Principal components of voxel-wise semantic models tile the cortex

a: Four principal components of the voxel-wise encoding model weights and associated words.
b: RGB color map corresponding to the first three principal components of the semantic space.

Later, Jain and Huth were pioneered the use contextual WE as brain encoding models, using activations from LSTMs (Hochreiter & Schmidhuber, 1997) to predict fMRI data, and showed that contextual WE allows for better predictions (compared to static WE), especially in high-level language areas (Jain & Huth, 2018). In parallel, Kell and colleagues (Kell et al., 2018) trained a modular convolutional neural network on two tasks: speech recognition and music genre identification. They found that, similar to the brain, the optimal network shares acoustic information early on, then splits into branches specific for each task. Unsurprisingly, the "word" branch correlated with the voxels in the language network, while the "genre" branch correlated better with voxels sensitive to music.

Toneva and Wehbe aligned many neural language models with fMRI data and discovered that in many cases the middle layers were the ones that achieved best performance (Toneva & Wehbe, 2019). They also found that enforcing uniform attention in the early layers improved the fit to the brain and, surprisingly, also the performance of the model on the task it was trained on. This suggests that aligning NLP models to brain data could be a way forward to boost their performance. It also corroborates other studies that show that the better a model is at the language modeling task, the better its brain score is (Schrimpf et al., 2021; Caucheteux & King, 2022), although the relationship might be more complex than a simple monotonic trend (Pasquiou et al., 2022).

In a series of studies using both fMRI and MEG, Caucheteux and colleagues also address the question using contextual WE derived from Transformers. They showed in a large-scale open fMRI dataset that the performance of the fit between model activations and the brain correlates with individual subjects comprehension scores, strongly suggesting that the fit is not due to low-level confounds (Caucheteux et al., 2021a). Crucially, they also separated neural language models activations into syntactic and semantic components, and found that purely syntactic representations aligned with a widely distributed network (Caucheteux et al., 2021b), arguing against the localist view of syntactic processing (Friederici, Opitz, et al., 2000)

Interestingly, the model's architecture, objective function, and training dataset all seems to influence its brain score (Schrimpf et al., 2021; Pasquiou et al., 2022).

Using intracranial grids of electrodes, Goldstein and colleagues found that the ordering of layers in a very large (48 layers) transformer matches well the temporal ordering of information processing in the brain's higher language regions (Goldstein, Ham, et al., 2022). Specifically, the early layers' brain score peaked around word onset, whereas for the later layers it was a few hundred milliseconds later. Crucially, this was not the case in lower-level language regions, where the peak in brain score happened at the same time for all layers. (Figure 0.12).

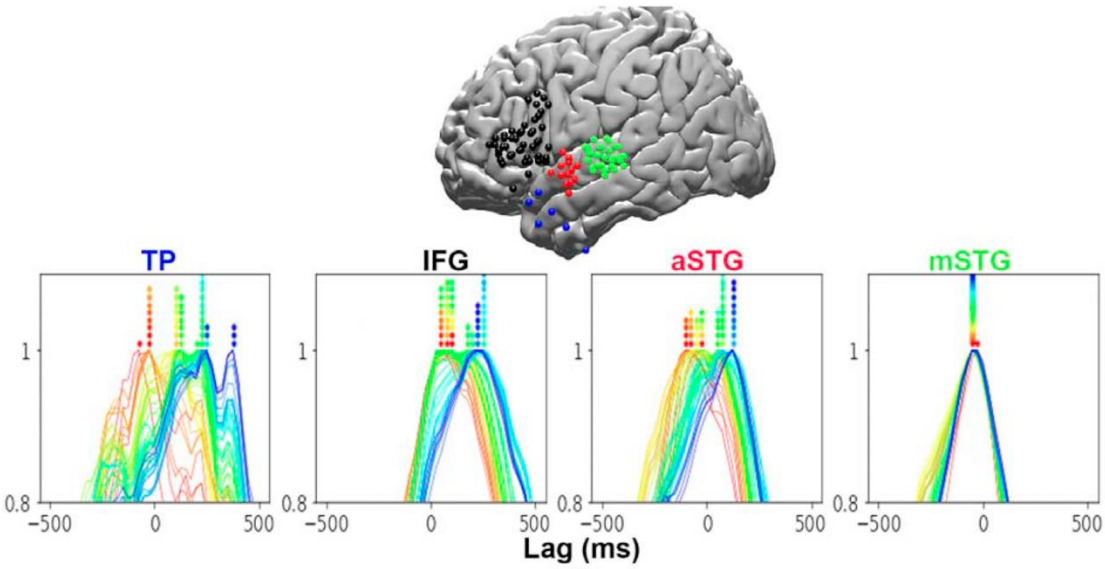


Figure 0-12: The layered hierarchy of large transformer language model maps to the temporal ordering of information processing in high level language areas

Top: electrode coverage for each region of interest: temporal pole (TP, blue), inferior frontal gyrus (IFG, black), anterior superior temporal gyrus (aSTG, red), and medial superior temporal gyrus (mSTG, green).
Bottom: normalized encoding performances in each of these regions, for each layer in the Transformer model, colored from red for early layers, to green for middle layers, and finally blue for higher layers.

In another pioneering study, Goldstein and colleagues showed that densely recorded activity from the IFG behaved similarly to contextual word embeddings even in a strict zeroshot mapping setup (Goldstein, Dabush, et al., 2022a). In other words, no encoding model was needed to find similarities in the geometries of brains' and models' embeddings, suggesting stronger similarities than previously thought.

Moving out of text and toward speech processing, these results were confirmed in a recent study (Millet et al., 2022) that showed that a self-supervised speech recognition model, Wav2Vec 2.0 (Baevski et al., 2020), could reach brain-like representation to speech from an unlabeled speech dataset of a size similar to what an infant would get to learn a language. Specifically, a specialization similar to the one found in the human cortex was identified, with sound-generic, speech-specific and language-specific representations.

Overall, despite considerable implementational differences, contextual word embeddings derived from neural language models have emerged as the state-of-the-art features to predict brain activations to linguistic inputs. Thus, they provide a new valuable tool to study semantic networks in the brain. This parallels a similar trend in vision neuroscience, where artificial neural networks trained on computer vision are now considered the best predictive models of brain activations during visual tasks (Yamins & DiCarlo, 2016; Schrimpf, Kubilius, Lee, et al., 2020; Zhuang et al., 2021).

F. Literature summary

To sum-up this literature review, many studies confirmed a core prediction from linguistics: that language is hierarchically organized, and that the brain uses this property to parse linguistic input. Neural signatures of these operations were found in specialized brain regions (Frankland & Greene, 2015), neural oscillations (Ding et al., 2016), transient and ramping activations (Pallier et al., 2011; Nelson et al., 2017). Multiple brain regions seem to be implicated, with a chief role of the left ATL and IFG (Pylkkänen, 2019).

However, doubts still remain about the actual function of these signatures. What computations allow to build temporary links between lexical elements, giving rise to compositional representations? Can these computations be isolated in space? In time? What is the format of such representations?

G. What this thesis tries to tackle

In this thesis, I sought to tackle these questions using tools from neuroimaging, linguistics, and artificial intelligence. I report two studies that provide complementary views on semantic composition in the brain.

In the first chapter, I will describe a study that investigated the neural dynamics and geometry of semantic composition in a joint MEG and intracranial EEG dataset, as well as deep neural language models. We isolate semantic processes by comparing normal sentences to meaningless pseudowords. Starting from theoretical models of lexical,

compositional, and wrap-up processes, we use multivariate decoding to separate these three processes and describe an extended network of brain regions implicated in each stage. Additionally, in an attempt to achieve a better understanding of the nature of compositional representations, we introduce to the field a new intrinsic dimensionality measure and show that meaningful representations are associated with higher dimensionality.

In the second chapter, we complement these findings by dissecting the processing of individual words in phrases of increasing length that have to be remembered and matched with a subsequent image. We show that words waiting to be composed are maintained for a longer time in neural activity when they have to be combined with other words. In addition, we demonstrate that compositional representations in working memory are compressed, such that neural activity is strongly affected by a measure of the complexity of the sentence. Finally, we show that the read-out from working memory is dependent on this complexity and is structure dependent, such that it takes longer to access syntactically deeper properties. Thus, we dissected the different phases of compositions: online integration, working memory storage, and final read-out.

Overall, this work brings us one step closer to the characterization of compositional representations. This should prove to be a fertile ground for future neural theories of compositionality, as our results can be taken as new constraints on the implementation of merging operations, and the dynamics and format of compositional representations.

Introduction to chapter 1

This first study holds a particular sentimental value for me, as it was the first experiment that I conceived, together with Stanislas, back in 2016 during a summer internship. The data was acquired by collaborators in the Timone hospital in Marseille during the following years, a partnership which offered us the possibility to collect a rare combination of magnetoencephalography (MEG) and intracranial electroencephalography (EEG) recordings. I started analyzing the data at the beginning of my PhD, three years later.

For this study, I implemented the sentence generation and stimulus presentation scripts; I set up classical signal processing and event-related-potentials and fields analyses, as well as multivariate decoding methods, in order to characterize the temporal dynamics of compositional processes. In addition, I trained recurrent neural networks and transformer language models in French, to be used as a testbed for our hypotheses and applied the same analyses on their activations. We also introduced to the field a new intrinsic dimensionality measure as a first step towards the characterization of the format of compositional representations.

The analysis of intracranial data was particularly challenging, but definitely worth the effort, as the effects were found to be much stronger than in MEG.

This work has been submitted to the Journal of Neuroscience and was accepted in March 2023.

Noteworthy, the “Jabberwocky” stimuli (sentences made of pseudowords) used in this study are inspired by the famous poem by Lewis Carroll (Carroll, 1871), introduced into neuroscientific studies by (Hahne & Jescheniak, 2001a). In this piece, Lewis Carroll pioneered the use of nonsense words, i.e., pseudowords with very limited semantic content but clear morphosyntactic markers that allow the identification of the parts-of-speech and thematic roles. Here is the whole poem, for your enjoyment:

JABBERWOCKY

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

"Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!"

He took his vorpal sword in hand:
Long time the manxome foe he sought—
So rested he by the Tumtum tree,
And stood awhile in thought.

And as in uffish thought he stood,
The Jabberwock, with eyes of flame,
Came whiffling through the tulgey wood,
And burbled as it came!

One, two! One, two! And through and through
The vorpal blade went snicker-snack!
He left it dead, and with its head
He went galumphing back.

"And hast thou slain the Jabberwock?
Come to my arms, my beamish boy!
O frabjous day! Callooh! Callay!"
He chortled in his joy.

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

from Through the Looking-Glass, and What Alice Found There (Carroll, 1871)

Chapter 1. Dimensionality and ramping: Signatures of sentence integration in the dynamics of brains and deep language models

Théo Desbordes*, Meta AI Research, Paris, France & Cognitive Neuroimaging Unit NeuroSpin center 91191, Gif-sur-Yvette, France

Yair Lakretz, Cognitive Neuroimaging Unit NeuroSpin center 91191, Gif-sur-Yvette, France

Valérie Chanoine, Institute of Language, Communication and the Brain, Aix-en-Provence 13100, France & Aix-Marseille Université, CNRS, LPL, Aix-en-Provence 13100

Maxime Oquab, Meta AI Research, Paris, France

Jean-Michel Badier, Aix Marseille Univ, INSERM, INS, Inst Neurosci Syst, Marseille, France

Agnès Trébuchon, Aix Marseille Univ, INSERM, INS, Inst Neurosci Syst, Marseille, France & APHM, Timone hospital, Epileptology and Cerebral Rythmology, Marseille, France

Romain Carron, Aix Marseille Univ, INSERM, INS, Inst Neurosci Syst, Marseille, France & APHM, Timone hospital, Functional and Stereotactic Neurosurgery, Marseille, France

Christian-G. Bénar, Aix Marseille Univ, INSERM, INS, Inst Neurosci Syst, Marseille, France

Stanislas Dehaene, Université Paris Saclay, INSERM, CEA, Cognitive Neuroimaging Unit, NeuroSpin center, Saclay, France ; and Collège de France, PSL University, Paris, France

Jean-Rémi King, PSL University, CNRS & Meta AI Research, Paris, France

* Corresponding author

Abstract

A sentence is more than the sum of its words: its meaning depends on how they combine with one another. The brain mechanisms underlying such semantic composition remain poorly understood. To shed light on the neural vector code underlying semantic composition, we introduce two hypotheses: First, the intrinsic dimensionality of the space of neural representations should increase as a sentence unfolds, paralleling the growing complexity of its semantic representation, and second, this progressive integration should be reflected in ramping and sentence-final signals. To test these predictions, we designed a dataset of closely matched normal and Jabberwocky sentences (composed of meaningless pseudo words) and displayed them to deep language models and to 11 human participants (5 men and 6 women) monitored with simultaneous magnetoencephalography and intracranial electroencephalography. In both deep language models and electrophysiological data, we found that representational dimensionality was higher for meaningful sentences than Jabberwocky. Furthermore, multivariate decoding of normal versus Jabberwocky confirmed three dynamic patterns: (i) a phasic pattern following each word, peaking in temporal and parietal areas, (ii) a ramping pattern, characteristic of bilateral inferior and middle frontal gyri, and (iii) a sentence-final pattern in left superior frontal gyrus and right orbitofrontal cortex. These results provide a first glimpse into the neural geometry of semantic integration and constrain the search for a neural code of linguistic composition.

A. Significance statement

Starting from linguistic theory, we make two sets of predictions in neural signals evoked by reading multi-word sentences. First, the intrinsic dimensionality of the representation should grow with additional meaning. Second, the neural dynamics should exhibit signatures of encoding, maintaining, and resolving semantic composition. We successfully validated these hypotheses in deep Neural Language Models, artificial neural networks trained on text and performing very well on many Natural Language Processing tasks. Then, using a unique combination of magnetoencephalography and intracranial electrodes, we recorded high-resolution brain data from human participants while they read a controlled set of sentences. Time-resolved dimensionality analysis showed increasing dimensionality with meaning, and multivariate decoding allowed us to isolate the three dynamical patterns we had hypothesized.

B. Introduction

To understand a sentence, the human brain must link each word to its meaning and bind these successive representations into an integrated representation of the sentence. It not only requires the maintenance of word meanings over time but also, crucially, the use of their syntactic relationships (Chomsky, 1957; Friederici, 2011; Pallier et al., 2011; Dehaene et al., 2015; Fedorenko et al., 2016; Ding et al., 2016; Nelson et al., 2017; Hagoort, 2019; Russin et al., 2019; Fedorenko et al., 2020; Caucheteux et al., 2021b). The neural basis of such compositionality has been studied in the case of two words composition (Pylkkänen, 2019, 2020a), but remains largely unknown for longer constituents.

In the present paper, we propose and put to an empirical test a new idea on how the brain encodes word sequences. Applying the neural population framework (Georgopoulos et al., 1986; Maass et al., 2002; M. M. Churchland et al., 2012; Yuste, 2015; Ebitz & Hayden, 2021) to language processing, we hypothesize that the construction of meaningful representations will necessarily lead the brain to recruit an increasingly large vector subspace. More specifically, we argue that the size of the neural manifold should grow with the progressive addition of meaning over the course of a sentence and propose two sets of predictions which are general consequences of this vector framework.

First, for each word that the subject reads and integrates in a compositional representation, this framework predicts an increase in the *intrinsic* dimensionality of the corresponding neural representation, i.e. the number of independent dimensions that actually participate in the encoding of this composed structure (Carreira-Perpinán, 1997). The idea is that, within the large dimensionality of the overall neural space (equal to the number of relevant neurons), only a much smaller vector subspace is actually used for encoding. Intrinsic dimensionality is thus defined as "the dimensionality of the manifold that approximately embeds the data" (Del Giudice, 2021). We predict that, for sentences, intrinsic dimensionality would increase as new meaning elements are put together: when we combine real words with one another, we generate a meaning which is more than the sum of its parts, and thus requires additional dimensions. Intuitively, one can think of concept cells (Quiroga et al., 2005) being recruited to encode the meaning of each incoming element, but also their relationships to each other (e.g. subject, complement, etc). When

processing a meaningless sentence, these additional dimensions would not be recruited. Consider the case of a Jabberwocky sentence, where meaningless pseudowords replace actual words while preserving the overall syntactic structure of the sentence (e.g., 'The cat jumped on the mat' would become 'The tula risped on the plor' (Mazoyer et al., 1993; Hahne & Jescheniak, 2001b). We predict that such a pseudo-sentence would activate a reduced set of semantic dimensions, because ``tula'', ``risps'' and ``plor'' carry little information about their e.g. size, form, usage and relationships. Note that there are a number of alternative hypotheses regarding the relative intrinsic dimensionality of normal and Jabberwocky sentences. Predictive coding theories of language processing (Shain et al., 2020; Heilbron et al., 2022; Goldstein, Zada, et al., 2022) forecast increased brain activations to surprising words, which in our setup should lead to normal sentences having the lowest responses. Likewise, if brain activity relates to processing difficulty (Carpenter et al., 1999; Just et al., 1996), then Jabberwocky should lead to the highest response. Thus, the predicted increase in dimensionality, greater for normal than for Jabberwocky sentences, should be found if and where brain signals are dominated by compositional semantics.

Our second set of predictions relates to the dynamics of this change of intrinsic dimensionality. We predict that meaningful composition will lead to a growing superposition of neural codes each recruiting additional neural dimensions, and thus leading to ramping neural activity over the course of the sentence. Furthermore, the representations of *normal* sentences should increase more than *Jabberwocky* sentences', in spite of their identical syntactic structures.

Several studies support these predictions (Fedorenko et al., 2016; Nelson et al., 2017; Pallier et al., 2011). For example, ramping signals reflecting the formation of linguistic constituents have been observed during the processing of normal sentences, first through fMRI (Pallier et al., 2011) and then electro-encephalography (EEG), magneto-encephalography (MEG) and intracranial recordings, including broadband signals (Ding et al., 2016; Caucheteux & King, 2020; Burroughs et al., 2021), beta (Bastiaansen et al., 2009; A. G. Lewis et al., 2015) and high-gamma frequency bands (Fedorenko et al., 2016), with peaks at the end of syntactic constituents (Nelson et al., 2017). Such ramping brain activity is likely to at least partially reflect semantic composition, as it has been reported to be larger for normal sentences than for their Jabberwocky counterparts (Fedorenko et al., 2016).

However, these previous studies did not disentangle *multi-word integration* from other semantic processes, such as *lexical access* and sentence-final *wrap-up*. Here, we reasoned that all these processes should differ in normal and Jabberwocky sentences, and we used the tools of multivariate decoding with temporal generalization (Fyshe, 2020; King & Dehaene, 2014) to disentangle them (See Methods for details). Obviously, several levels of internal representation should allow a simple classifier to categorize the incoming stimulus as normal or Jabberwocky. However, the evolution of these representations over time should help identify specialized components. We derived a set of theoretical generalization matrices that depicts the expected dynamics of classifiers trained to separate normal and Jabberwocky sentences from neural activity, for each of the three processes under study (Figure 1D). These three stages are not tied to a particular theory of sentence processing (Just & Carpenter, 1980; Seidenberg & McClelland, 1989; Frazier & Clifton, 1996; Steedman, 2001; R. L. Lewis & Vasishth, 2005); rather, we propose them as necessary steps in sentence comprehension and study them as such.

First, *lexical access* triggered by each word is predicted to elicit a *phasic*, transient response that differs for normal and Jabberwocky words (Just & Carpenter, 1980; Seidenberg & McClelland, 1989; Caramazza, 1997) (Figure 1D red). This pattern is expected to be found in the anterior fusiform gyrus (FuG) and the superior and middle temporal gyri (STG, MTG), peaking between 250 and 400 ms after word onset (Nobre et al., 1994; Binder et al., 2003; Woolnough et al., 2020).

Second, as stated above, multi-word integration is predicted to elicit ramping dynamics, characterized by an increasingly strong square pattern in the temporal generalization matrix (Figure 1D blue), especially in the Inferior Frontal Gyrus (IFG; (Fedorenko et al., 2016; Nelson et al., 2017; Pallier et al., 2011)).

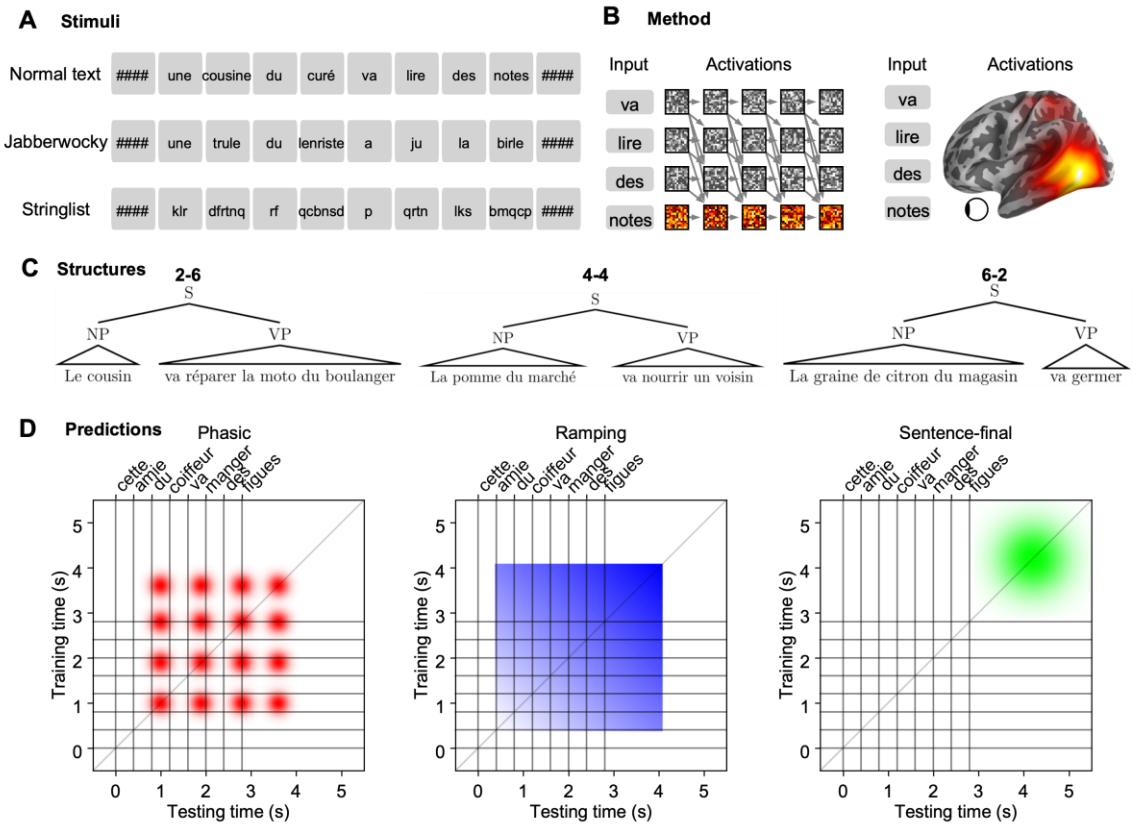


Figure 1-1: Experimental design and template matrices

A. Example stimuli for the 3 conditions: normal text, jabberwocky, and list of consonant strings (stringlist). Masks of '#' were presented before and after each sentence in order to keep visual masking approximately constant.
B. Extraction of activations from Neural Language Models and combined MEG-sEEG brain recordings.
C. Example of the 3 syntactic structures used in this study, varying in the relative size of the NP and the VP.
D. Theoretical temporal generalization patterns. Regions involved in lexical processing of single words would differentiate normal content words from jabberwocky content words based on lexical access mechanisms (the function words are the same in both conditions), yielding a phasic pattern. Regions involved in compositional processes should exhibit a ramping generalization pattern, where normal and jabberwocky sentences become more and more differentiated with each incoming word. Finally, regions involved in wrap-up processes would separate normal and jabberwocky only after the sentence is finished, leading to a sentence-final pattern.

Supplementary Figure 1 shows electrodes coverage.

Last, sentence-final *wrap-up* processes are predicted to take place following the onset of the last word, and as such elicit a *sentence-final* distinction of normal and Jabberwocky sentences (Figure 1D green). Indeed, readers are known to pause at the end of sentences to integrate, interpret, and incorporate the constituting elements into the general context of the discourse (Just & Carpenter, 1980). Behavioral studies based on eye-tracking

have evidenced an increased reading time for sentence-final words (Warren et al., 2009; Kuperman et al., 2010). Furthermore, end-of-sentence effects reflecting in-sentence grammatical gender violation have been observed in EEG (Molinaro et al., 2008), and intracranial EEG signals have been observed peaking at sentence ending (Nelson et al., 2017). An fMRI study found that markers of syntactic complexity were absent during sentence processing, but appeared when the subjects were probed to extract structural information not obvious in the superficial sequence of words (Pattamadilok et al., 2016). Finally, it was found that sentences with a negation were marked as such in working memory and necessitated an increased processing time, even after a long delay (Agmon et al., 2022). As such, this sentence-final difference in representations is expected to differ, at least partially, from that present during the sentence (as seen in the lack of generalization to earlier time points in Figure 1D green).

Here, we test the above predictions in both humans and artificial neural networks. Neural networks have long been used to model natural language processing (Rumelhart & McClelland, 1986; Pinker & Prince, 1988; Sharkey, 1992; Pater, 2019; Oh et al., 2022), and neural language models (NLMs) trained on next-word prediction have recently undergone a revival as models of human language processing (McClelland et al., 2020; Hale et al., 2021) and learning (Warstadt & Bowman, 2022) (although also see (Lakretz, Desbordes, Hupkes, et al., 2021)). Here we use NLMs as a testbed to check whether our hypotheses can be verified in a noise-free language processing system that has major implementational differences compared to biological neural networks. Convergent findings would support the view that we are studying general properties of linguistic composition. Several researchers have started to analyze activations from NLMs in order to attempt to shed light on the neural codes for language (Tenney et al., 2019; Clark et al., 2019; Lakretz et al., 2019; Rogers et al., 2020), and the present work contributes to this field by introducing the temporal generalization method.

In brief, the present work aimed to address three main questions: how does the dimensionality of the neural representation evolve during sentence processing? Can phasic, ramping and sentence-final signals be disentangled in brain dynamics? Do they occur in separate brain regions? We tested our hypotheses first in NLMs, including home-trained character-based Transformers (Vaswani et al., 2017) and LSTMs (Hochreiter & Schmidhuber,

1997), as well as CamemBERT (L. Martin et al., 2020), then in electrophysiological recordings of 11 human participants whose brain activity were simultaneously recorded with magneto-encephalography (MEG, n=276 sensors per subject) and intracranial stereotactic electro-encephalography (sEEG, n=2,243 electrodes in total, see electrode placement in Supplementary Figure 1).

We start with a quick overview of the diversity of neural signals in our dataset. We then present the intrinsic dimensionality analysis, followed by multivariate decoding in NLMs and brains. Finally, we quantify the presence of each theoretical pattern in the empirical generalization matrices by means of multiple regression and replicate the decoding analysis in multiple brain regions. Overall, our results back the idea that learning language is associated with a predictable shaping of the representational manifold, such that meaningful representations call upon a larger number of representational dimensions and promote the assignment of neural dimensions to meaning.

C. Methods

1. Ethics

Eleven right-handed individuals (5 men and 6 women; age range=25–57, mean= 40, SD= 9.4) with intracranial stereotactic electrodes implantation as part of their treatment for refractory epilepsy gave their informed consent to participate in our study, in accordance with the ethic evaluation RCB 2018-A02363-52. All patients were implanted with depth electrodes for clinical purposes (presurgical evaluation) in the Epileptology and Cerebral Rythmology Department of the Timone Hospital (Marseille, France). Neuropsychological assessment indicated that all patients had intact language functions. Their reading ability was controlled by means of a French version of a reading test (test Malabi, © Unité INSERM-CEA de Neuroimagerie Cognitive).

2. Stimuli and task

Sentences of 8 words were presented to the participants in a Rapid Stream Visual Presentation with an SOA of 400ms. Each sentence was preceded and followed by visual masks (####) in order to keep forward and backward masking constant (Figure 1A).

The stimuli were generated using a custom sentence generator script that constructs a wide range of sentences from a finite vocabulary set, respecting several constraints:

- The sentences were 8 words long.
- Each sentence comprised a systematic alternation of short function words (determiners and auxiliary) and longer content words (nouns and verbs).
- The sentences consisted of a Noun Phrase (NP) followed by a Verb Phrase (VP). The NP consisted of a determiner, a noun (the subject of the sentence) and optionally one or two prepositions. The VP consisted of an auxiliary, a verb and optionally a determiner and a noun (the object of the verb) and one or two prepositions.
- There were 3 possible syntactic structures that varied in the size of the NP and VP. They could both be of size 2, 4, or 6 words, while their sum was always equal to 8 words (Figure 1C).

The vocabulary consisted of:

- 9 determiners, i.e., 3 for each gender in the singular form and 3 for the plural form.
- 10 verbs that could appear in singular or plural, in the present or past tense (40 different forms).
- 75 nouns, among those 46 could appear in the singular or plural form (the others were always singular), 46 were masculine (26 were used as subjects and objects, the rest appeared in prepositional phrases), 26 were feminine (all were used as subjects and objects) and 3 could appear in either masculine or feminine form (used as subjects only).

The total number of distinct sentences the script could generate was 791,754. For each subject, in the normal and jabberwocky conditions, we sampled an equal number of

each syntactic structure, as well as an equal number of feminine/masculine and singular/plural subjects and objects and an equal number of present/past tense for the verb.

The Jabberwocky stimuli were designed by hand, by changing one or two letters to create nonwords but keeping the morphological markers present.

Strings lists consisted of strings of consonants of similar length to the actual words. These letter strings did not have any morphosyntactic information and thus constituted a low-level, mainly visual, control to the linguistic stimuli used in the experiment.

The task was for the participant to detect the presence of target words in the sentences and press the response button as fast as possible when the target was present. The target was present in 1/11 sentences.

Participants performed 330 trials in total (cut down in 6 blocks), composed as follows:

- 120 normal sentences
- 120 jabberwocky sentences
- 60 strings of consonants
- 30 sentences of a random condition containing the target words and that are discarded from the analyses.

3. sEEG and MEG data acquisition and preprocessing

MEG and SEEG recordings took place simultaneously in a dimly illuminated, magnetically shielded room. Recordings were obtained from subjects in supine position to limit the movement during the recording.

MEG signals were acquired with a 248-channel biomagnetometer system (Magnetometers. 4D Neuroimaging, San Diego, CA, USA located in the MEG facility, Timone Hospital, Marseille). The data were recorded continuously with a band- DC-800 Hz bandwidth with a sampling rate of 2034.51 Hz.

SEEG recordings were performed using intracerebral multiple contact electrodes (10–15 contacts, length: 2 mm, diameter: 0.8mm, 1.5 mm apart from edge to edge) placed

intracranially according to Talairach's stereotactic method (Bancaud et al., 1970; Talairach et al., 1992).

SEEG as well as EOG and ECG signals, to facilitate the ulterior rejection of eye movements, blinks, and cardiac artifacts, were simultaneously recorded with MEG (Badier et al., 2017) with a band- 0.01-1000 Hz bandwidth with a sampling rate of 2500 Hz using a 256-channel BrainAmp amplifier system (Brain Products GmbH, Munich, Germany). SEEG was then interpolated at the sampling rate of the MEG thanks to triggers.

In order to determine the location of the head with respect to the MEG helmet, five coils were fixed on the subject's head. The position of these coils as well as the surface of the head were digitized with a 3-D digitizer (Polhemus Fastrack, Polhemus Corporation, Colchester, VT, USA), and head position was measured at the beginning and at the end of each run. The head shape obtained from the digitization of the head was used to check and eventually compensate for differences in head position between runs or to match to the participant's MRI. All stimuli were presented to the subjects on a mirror by a back-projection system where an LCD projector was placed outside the magnetically shielded room in order to avoid interfering electrical apparatus. The distance between the participant's eyes and the screen on which stimuli were displayed was similar across patients. A trigger square invisible to the participant was projected onto a photodiode which was used to signal the presence of a stimulus on-screen and to synchronize the MEG and EOG/ECG recordings.

Among the eleven patients, nine underwent an MEG recording at the same time.

The sEEG and MEG data were band-pass filtered at 0.3-500hz, notch filtered at 50h and the first 3 harmonics to remove line noise. The data were then downsampled to 100hz and clipped at 10 times the standard deviation either side of the median value, separately for each channel. We then used an automatic detection procedure for bad channels in which the temporal variance is computed for each channel and a value above or below 25 times the median variance over channels leads to rejection.

Epochs were constructed keeping time points from -0.5 to 5.5 seconds after the onset of the first mask. We then used a procedure to reject bad epochs similar to the one

we used for channels: the variance was now computed over time and the remaining channels, separately for each epoch, and a value above or below 5 times the median over epochs lead to rejection. Baseline correction was then applied using the 400ms interval between the onsets of the first mask and the first word. The data were then smoothed using a 100ms hanning window. Finally, Common Median Referencing was applied using all channels but the ones that were marked as bad. All the data preprocessing was done using the MNE-Python software (Gramfort et al., 2013).

The high-gamma activity (in Figure 2) was extracted using a Morlet transform and 8 frequency bands linearly spaced between 70 and 150 hertz, then combined with Principal Component Analysis, which we found to be more robust than simple averaging. For all subsequent analyses we used raw voltage instead high-gamma because we found the overall decoding performance to be better. However, we replicated the decoding and dimensionality results with high-gamma and did not find major differences.

4. Localizer test

To boost the statistical power, we selected language-specific electrodes using a two-sample temporal cluster permutation test. Specifically, we tested whether the conditions (normal text, jabberwocky) and (stringlist) were different at the whole epochs level. We kept electrodes that contained at least one cluster after the permutation test and FDR correction. We used 1000 permutations and threshold-free cluster enhancement (S. M. Smith & Nichols, 2009) with a starting threshold of 0 and a step of 0.1.

5. Dimensionality Analysis

We used a previously reported method based on Principal Component Analysis (PCA) in order to compute the intrinsic dimensionality (ID) (Elmoznino & Bonner, 2022; Gao et al., 2017), sometimes called the participation ratio (Sorscher et al., 2021). This method quantifies intrinsic dimensionality as follows:

$$D = \frac{(\sum_{i=1}^{M} \lambda^i)^2}{\sum_{i=1}^{M} \lambda^{i^2}}$$

where λ^i are the eigenvalues of the neural covariance matrix (i.e., the eigenvalues whose corresponding eigenvectors are the principal components of the dataset), and M is the number of channels (electrodes or magnetometers). This gives a continuous measure of the number of principal components needed to explain most of the variance in a dataset. Intuitively, one can check that if the data varies only along a single dimension, all of the variance will be explained by the first principal component, hence a single eigenvalue λ^1 will be non-zero, and therefore the formula implies that D = 1. On the contrary, if the signals in all channels vary independently of each other (and with similar magnitude), such that each principal component explains an equal part of the variance, then D will be equal to the number of channels. Between those two extremes, D estimates the approximate number of dimensions that vary significantly in the brain signals.

To calculate D, we computed a PCA on 0.4 s sliding time windows, combining time points and trials (such that the PCA's input is a n_times*n_trials by n_channels matrix), separately for each of the 3 conditions (normal, Jabber, stringlist), and computed the ID of the resulting eigenspectra using the aforementioned formula. This analysis was repeated 10 times with different (non-overlapping) parts of the data to get an average D and the corresponding standard error bars.

6. Multivariate decoding

We trained a logistic regression to separate normal and jabberwocky sentences at each time point using MEG and sEEG single-trial data. Such a decoding analysis informs us about whether and when our two conditions are differently represented in neural signals: if at time *t* the classifier reaches above-chance performance, it means that the brain (or the specific region of interest) segregates normal and Jabberwocky stimuli at this time. These classifiers were then tested at each other time point according to the temporal generalization method (King & Dehaene, 2014). This extension of the traditional within-time decoding analysis allows to test for the consistency of neural patterns over time: if a classifier trained at time *t* generalizes to time *T*, it means that the neural patterns is somewhat similar between time *t* and *T*. On the other hand, within-time decoding could be high at both *t* and *T*, but with no generalization between *t* and *T*. This would mean that the brain segregates normal and Jabberwocky stimuli at both time points, but with a different

pattern of activations. In other words, the within-time decoding performance (trained and tested at the same time, i.e., the diagonal of the temporal generalization matrix) inform us about the content of brain signals, while the across-time decoding performance (trained and test at different times, i.e., the off-diagonal elements) tells us about the stability of these representations.

Before training the classifiers, the data was subtracted from its median and scaled using the interquartile range, i.e. the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile). We used a stratified k-fold crossvalidation procedure with 10 folds. We average the classifiers' performances across all splits and report the average performance across subjects. We also stored the AUC for 100 permutations of the test labels in order to assess significance of the regional pattern regression analysis. To infer the overall tendency to generalize, we averaged each line of the temporal generalization matrix.

These analyses were performed with scikit-learn (Pedregosa et al., 2011).

7. Template regression

We trained a linear regression to predict empirical AUC matrices (averaged over patients) from templates. Specifically, each 601x601 matrix (we have 601 time points) was flattened to a 601^2 vector, used in the regression analysis. For each template, we made a grid search to select the best parameters (delay and width, see Supplementary Figure 9). We thus tested, for each empirical matrix, 100 candidates for each kind of template (phasic, ramping, sentence-final). The best template was the one with the highest likelihood in the regression model.

8. Regional pattern analysis

Region of interests (ROI) were extracted from the Harvard- Oxford Cortical Atlas (Desikan et al., 2006). The whole decoding and pattern regression pipeline was repeated for each region. We only considered regions where at least three subjects had electrodes. The p values reported in this section have been FDR corrected at the region level.

9. Neural Language Models

We report the results of home-trained character-based Transformers and Long-Short Term Memory (LSTM, (Hochreiter & Schmidhuber, 1997) models and of CamemBERT (L. Martin et al., 2020), a BERT (Devlin et al., 2019) model trained on a very large French dataset. The activations extracted from the CamemBERT models were obtained by giving the sentence piece by piece to the model and averaging the activation of each wordpiece composing a word. Thus, although the model is bidirectional we only gave it information about the past up to the current word.

The character-based LSTM models had 2 layers of 1024 units, while the character-based Transformer models had 12 layers and 768 units per layer. Both were trained on a 2GB sample of the French Wikipedia (Merity et al., 2016). We trained the models for 20 epochs, an initial learning rate of 20, a batch size of 128, a dropout rate of 0.2, and a sequence length of 35. At the end of an epoch, if no improvement was seen on the validation set, the learning rate was halved. For both LSTMs and Transformers models we used 10 instantiations of the model with different random seeds and report the performance averaged over all seeds. To obtain a single activation vector per word, we used the average activation evoked by each character belonging to the word for the character embedding layer, and the activations at the last character of the word for each upper layer. Untrained models were initialized with random weights (all sampled uniformly between -0.1 and 0.1), and directly underwent the same procedure for extracting their activations. We chose character-based models because word-based models cannot generalize to Jabberwocky (they only take trained words as input), whereas character-based models can take any string as input. All model manipulations were done with Pytorch (Paszke et al., 2019)

For the decoding analysis, we used a sample of 1000 sentences of each condition, generated using the same script as the subjects.

D. Results

Eleven sEEG patients implanted for clinical purposes read 240 sentences in rapid stream visual presentation (RSVP) with an SOA of 400 ms. Among them, half were normal French sentences, and the other half were the syntactically matched Jabberwocky sentences. Each sentence consists of eight words alternating between function words (determiners and auxiliary in position 1, 3, 5 and 7) and content words (nouns and verbs or their equivalent Jabberwocky pseudowords in position 2, 4, 6 and 8). These stimuli were mixed with 60 "string lists" of similar length, which consist of meaningless sequences of strings (Figure 1A) and were used as a low-level control. Specifically, as a localizer test, before all analyses, we selected language-selective channels using temporal cluster permutation test: only the channels with at least one significant cluster - after False Discovery Rate (FDR) correction at the electrode level - when comparing i) normal or Jabberwocky sentences to ii) string lists were kept for subsequent analyses.

1. Diversity of sEEG responses during sentence processing

We begin with a quick descriptive overview of the diversity of brain signals across electrodes and patients. Figure 2 illustrates some of these responses, both in the evoked broadband domain and in the high gamma frequency range (> 70 hertz). In each electrode, we assessed the variation of brain responses with our experimental conditions using a temporal cluster permutation test. Electrodes with at least one significant cluster - after False Discovery Rate (FDR) correction at the electrode level – when comparing i) normal and ii) Jabberwocky sentences, for either broadband or high gamma signals, were considered. We then manually selected representative electrodes from this pool. We do not claim these results to be exhaustive, rather we find it helpful to hold in mind these illustrative neural signals when examining the subsequent analyses. Unless specified otherwise, all subsequent statistical tests in this section are two-sample Wilcoxon-Mann-Whitney tests.

First, several visual electrodes exhibited a fast phasic broadband response, triggered indifferently by all visual stimuli, but modulated by stimulus length. As previously reported in other datasets (Agrawal et al., 2020; King et al., 2020; Woolnough et al., 2020), short function words triggered a smaller response than longer content words ($p<0.01$ based on

the average activity between 50 ms and 300 ms following each stimulus presentation, electrode E6142, Supplementary Figure 2A). Note that this channel is the only exception to the selection rule stated above: its activity did not differentiate between normal and Jabberwocky sentences, neither in broadband nor high gamma power.

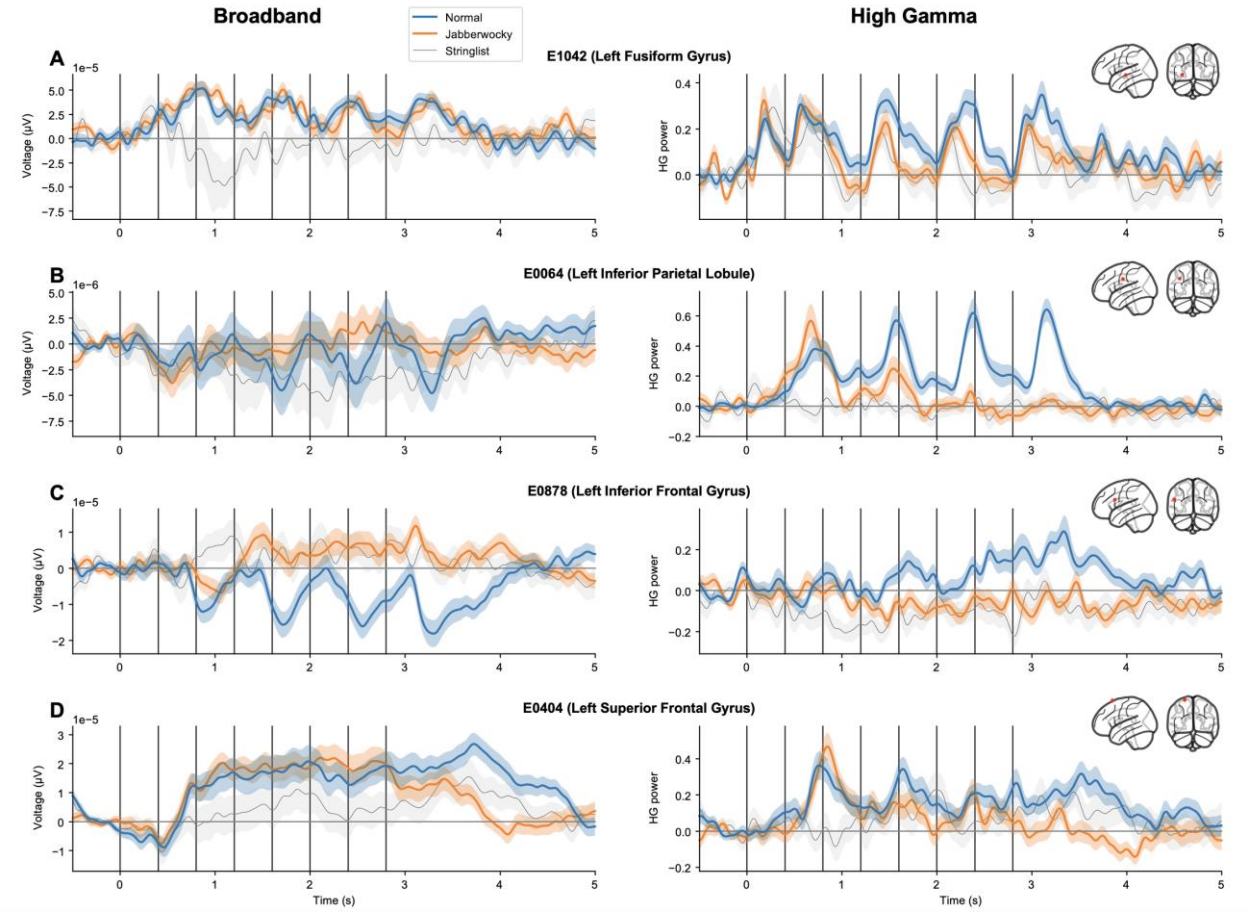


Figure 1-2: Illustrative profiles of human sEEG responses compatible with the postulated phasic, ramping and sentence-final patterns

Four examples of electrodes responding to normal sentences (blue), Jabberwocky sentences (orange), and string of consonants (gray). Each line shows the local field potential (Voltage, Left) and the high-gamma (HG power, Right) for the same electrode. The 8 vertical lines represent the onset of each word in the sentence.
Supplementary Figure 2 shows additional illustrative profiles.
Supplementary Figure 3 shows single units from a Transformer NLM.

Second, in the fusiform gyrus (FuG), the evoked responses of electrode E1042 (Figure 2A) was similar between normal and Jabberwocky, but significantly different from string lists ($p<0.0001$), consistent with this region's sensitivity to written words and word-like stimuli

(Woolnough et al., 2020). Interestingly, high gamma power from the same electrode exhibited phasic responses that differed for normal and jabberwocky ($p<0.01$, Figure 2A right), compatible with a role in lexical access (Woolnough et al., 2020).

Third, in regions such as the inferior parietal lobule (IPL, Figure 2B), high gamma responses to normal sentences showed a clear phasic effect. For example, in electrode E0064 it peaks 300ms after each content word. In this electrode, responses to Jabberwocky were very similar for the first pseudoword, but then quickly dropped.

While most electrodes exhibited stronger responses to normal sentences than Jabberwocky sentences, some responded specifically to Jabberwocky (e.g. superior temporal electrode E1263, Supplementary Figure 2B). Such differences are compatible with the dual-route model of reading (Marshall & Newcombe, 1973; Jobard et al., 2003; Coltheart, 2005): while words evoke additional lexical, syntactic, semantic and, ultimately, compositional processes, pseudowords may also elicit specific processes associated for instance with attention and grapheme-phoneme conversion (Rumsey et al., 1997; Binder et al., 2003; Taylor et al., 2013).

Fourth, in inferior frontal gyrus (IFG, Figure 2C) and medial frontal gyrus (MFG, Supplementary Figure 2C), we observed ramping responses: each additional content word led to an increase of the broadband and high gamma responses (linear regression on the difference in HG activity between normal and Jabberwocky for electrode E0878 (IFG) from 0 s to 3.5 s: slope=0.086, $r=0.78$, $p<0.0001$; and for electrodes E3652 (MFG) on the broadband: slope=$4.99e^{-6}$, $r=0.64$, $p<0.0001$). This is compatible with a role in linguistic integration, in line with previous studies (Fedorenko et al., 2016; Nelson et al., 2017).

Last, in left superior frontal gyrus (SFG, Figure 2D) and right orbitofrontal cortex (OFC, Supplementary Figure 2D), we observed sentence-final effects: for example, electrode E6062 (OFC), was mostly silent during the sentence, started to increase toward its ending, and exhibited a sharp peak more than 1 s after the last word's onset (Supplementary Figure 2D). On the other hand, electrode E0404 (SFG) responded similarly to both normal and Jabberwocky words, but these conditions ultimately diverged after the last word (Figure 2D).

Overall, the broad spectrum of functional responses illustrates the difficulty of interpreting the neural bases of language. Interestingly, a similar diversity of responses can be observed in individual units of deep language models (Supplementary Figure 3). In the following sections, we use multivariate dimension reduction and decoding tools to evaluate whether our theoretical framework can account for the latent structure underlying these complex neural signals.

2. Evaluating the intrinsic dimensionality hypothesis

As detailed in the introduction, our framework predicts that string lists, Jabberwocky sentences and normal sentences should lead to neural representations that systematically increase in their intrinsic dimensionality (ID). To compute the ID, we follow previous studies in neuroscience outside of the language domain (Elmoznino & Bonner, 2022; Gao et al., 2017; Sorscher et al., 2021) and use a method based on Principal Component Analysis (see Methods).

We performed this analysis on 400 ms time windows from -0.4 s to 5.2 s. Figure 3 shows the dimensionality estimate as a function of window onset. Three findings fit with our predictions. First, in all conditions for sEEG and for the normal text condition in MEG, the estimate increased with window onset, thus showing that as the successive words unfolded, the dimensionality of the brain signals increased (Pearson correlation with time, normal: $R=0.98$, Jabber: $R=0.94$; stringlist: $R=0.89$, $p<0.0001$ for each for broadband sEEG, Figure 3A; normal: $R=0.91$, $p<0.0001$, Jabber: $R=0.50$, $p>0.05$, stringlist: $R=0.47$, $p>0.05$ for MEG, Figure 3B, FDR corrected). Second, the intrinsic dimensionality was overall larger for normal sentences than for Jabberwocky sentences and string lists in the sEEG signals ($p<0.001$ for normal versus Jabber, $p<0.001$ for normal versus string lists, Wilcoxon-Mann-Whitney using the 4 s, FDR corrected, Figure 3A right). This was also the case for the MEG ($p<0.001$ for normal versus Jabber, $p<0.001$ for normal versus string lists, FDR corrected, Figure 3B right). Third, the difference increased as the sentence unfolded, as determined by a significant Pearson correlation between the difference (Normal – Jabber) and time ($R=0.96$, $p<0.0001$ for sEEG; $R=0.92$, $p<0.0001$ for MEG. We replicated these results with sEEG high gamma power and found highly similar results (Supplementary Figure 4A).

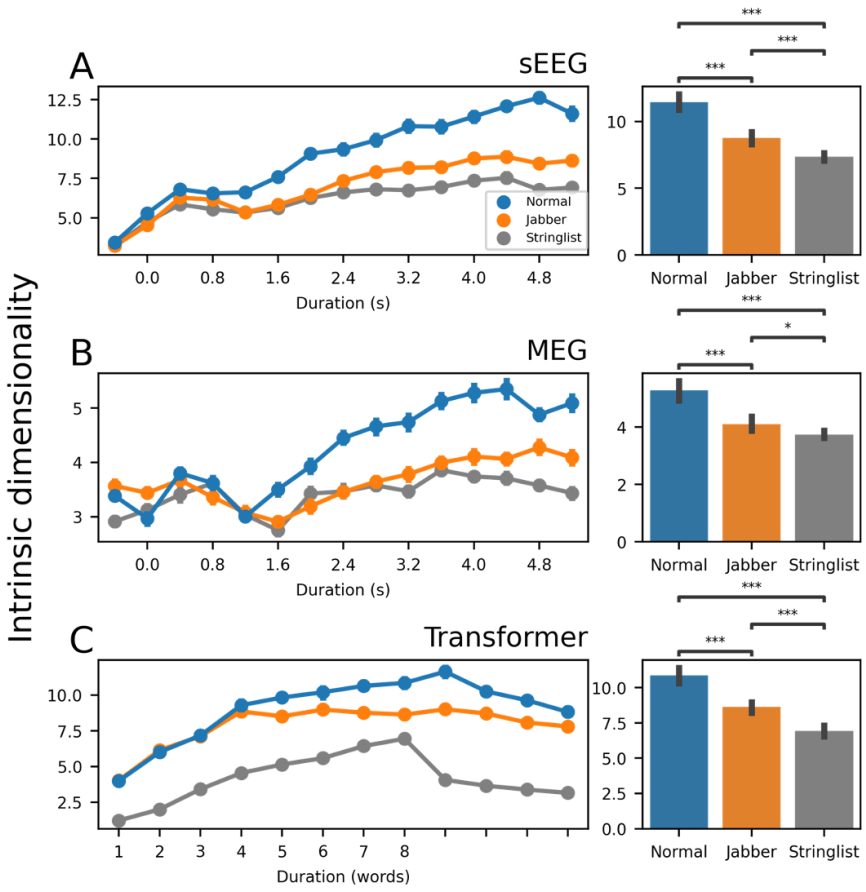


Figure 1-3: Intrinsic dimensionality is higher for normal sentences than Jabberwocky

A: Intrinsic dimensionality computed from the broadband sEEG signals from all subjects as a function of the time window used (sliding time window of 0.4 s width). Right: Bar plot showing the intrinsic dimensionality computed using the whole sentence (a full 4 s time window).

B, C, same analysis applied to MEG signals and Transformer activations (last layer, averaged over all stimuli for each condition). For the Transformer, the bar plot shows the intrinsic dimensionality computed using the 8 words time window.

Supplementary Figure 4 shows intrinsic dimensionality for HG sEEG, Camembert and LSTM.

Supplementary Figure 5 shows the lack of effect in untrained models.

Comparable effects were also observed in NLMs, such as causal Transformers (Figure 3C; $p<0.001$ for both pairwise comparisons; $p<0.001$ for all correlations), LSTMs (Supplementary Figure 4B; $p<0.001$ for both pairwise comparisons; $p<0.001$ for all correlation) and CamemBERT (Supplementary Figure 4C; $p<0.05$ for normal versus Jabber, $p<0.001$ for normal versus string list; $p<0.0001$ for all correlations). Crucially, untrained NLMs did not exhibit any significant differences between normal and Jabberwocky (Supplementary Figure 5). Thus, in models, language learning is associated with a reshaping of the representational manifold, with an attribution of meaning to specific dimensions.

3. Neural language models (NLMs) exhibit phasic, ramping and sentence-final responses

We next tested our second prediction, i.e. the existence of distinct phasic, ramping and sentence-final responses to sentences. To put these predictions to a test, we first evaluated whether these putative stages of semantic composition could be identified in NLMs using multivariate decoding and generalization. For this, we extracted the activations of each model in response to our stimuli and trained, at each word relative to sentence onset, a logistic regression across its artificial neurons to classify normal versus Jabberwocky sentences. We then evaluated these logistic regressions at each other time point, including the representations after the end of the sentence (obtained by feeding the model with four additional "space" tokens and extracting the corresponding activations). This temporal generalization analysis (King & Dehaene, 2014) resulted in a 12x12 training x testing matrix of classification scores summarizing i) where and when information distinguishing normal and Jabberwocky sentences is linearly represented in the network, and ii) whether the underlying representations change as the sentence unfolds.

The first computational step in NLMs is an "embedding" layer, where each vocabulary item (i.e., each character) is mapped onto a unique vector. Consequently, we expected temporal generalization to reveal a lexical signature in this embedding layer: i.e., a transient and phasic score rising after content words (Figure 1D). Our analysis confirmed this prediction: decoding leads to a relatively small above-chance decoding performance at each content word (Figure 4A, mean AUC over all content words: 0.79, Wilcoxon-Mann-Whitney test against chance: $p<0.0001$). As expected, this signature disappeared for function words, as they were identical for normal and Jabberwocky sentences.

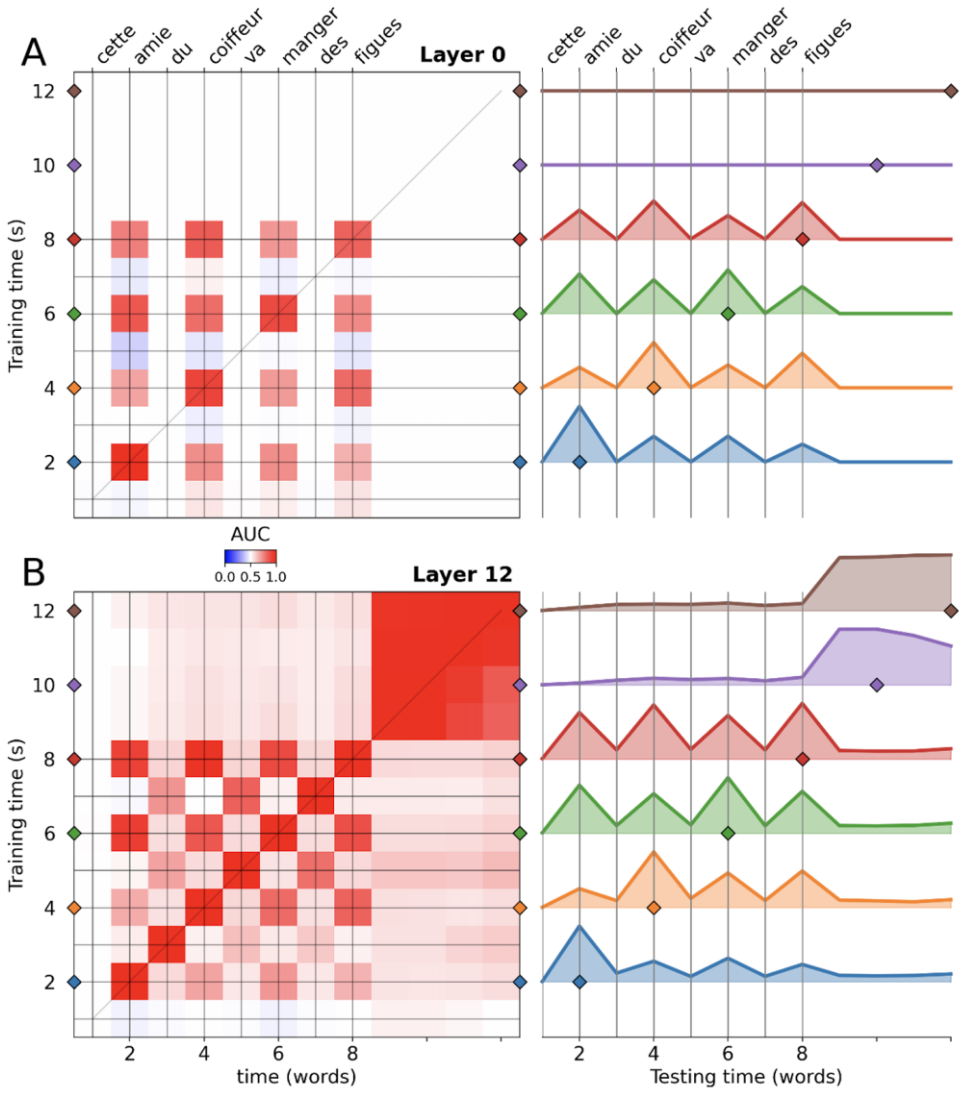


Figure 1-4: Decoding normal versus Jabberwocky sentences in neural language models shows lexical, persistent, and ramping patterns

Left: Temporal generalization matrices for a decoder trained to distinguish normal sentences from jabberwocky using the activity of the input word layer (top) and the final (12th) layer in a Transformer language model. The area under the curve (AUC) is the average over the 10 models trained on the same corpus but instantiated with different random seeds. Note that, in non-contextualized word embeddings, we only see the lexical pattern, whereas contextualized layers exhibit a superposition of multiple theoretical patterns. Although the performance on the diagonal is at ceiling (AUC=1), the generalization pattern is consistent with the ramping model.

Right: Generalization of individual decoders, i.e., horizontal slices from the temporal generalization matrices on the left. These slices from each matrix show in more details the temporal dynamics of sentence processing. Filled lines show significant time points, tested with Wilcoxon-Mann-Whitney test against chance (0.5) and FDR correction.

The first time point corresponds to the onset of the visual mask preceding the sentence. The onsets of successive words are marked by vertical grey lines. The small colored diamonds show the time where the decoders shown were trained.

Supplementary Figure 6 shows decoding performance for LSTM and Camembert.
Supplementary Figure 7 shows ramping slope as a function of layer for LSTM and Transformers.

The deep layers of NLMs integrate information from multiple tokens. Because of this integration of the preceding context, we expected temporal generalization to reveal an increasingly strong "square" of decoding performance. Indeed, we observed a temporal generalization not just between content words but also for function words as well as after the sentence. This was true for causal Transformers (Figure 4B), LSTMs (Supplementary Figure 6A) and CamemBERT (Supplementary Figure 6B). We show results for the last layer, where we found the overall decoding performance to be the strongest, nevertheless the earlier layers exhibited similar dynamics (Supplementary Figure 7). Because the diagonal - performance was at ceiling (AUC = 1) after the first content word, we could not directly test whether temporal generalization significantly increases over the course of the sentence, as predicted by a ramping processing stage (Figure 1D). However, the generalization performance of individual decoders (i.e., the average of each line from the matrix) increased over the course of the sentence, as revealed by a linear regression on the average of each line from word 1 to word 8 for layer 6, slope=0.017, r=0.78, $p<0.0001$). This finding suggests that NLMs demonstrate a ramping activity pattern that varies with semantic composition. Furthermore, this tendency to ramping increased in the upper layers of LSTM and causal Transformer models (Supplementary Figure 7), suggesting that higher-level linguistic information is characterized by a stronger ramping signature.

Finally, looking at activations after the end of the sentence (i.e. when NLMs are input with spaces, upper right in Figure 4B), we observed a strong square generalization pattern (mean AUC for models trained and tested on tokens 9 to 12 : 0.97; Wilcoxon-Mann-Whitney test against chance: $p<0.0001$) that generalizes only modestly to the preceding words (mean AUC of classifiers trained on tokens 9 to 12, generalized to all preceding words: 0.59; Wilcoxon-Mann-Whitney test against chance: $p<0.0001$). Consistent with the predictions of a wrap-up processing stage, this result suggests that sentence-final representations partially differ from those generated during online sentence processing.

Overall, these analyses confirm that temporal generalization can isolate the three putative processing stages of semantic composition in NLMs: a phasic effect at the earliest processing stage of the network, and ramping and end-of-sentence effects at higher

processing levels. In the next section, we apply these analyses on the sEEG and MEG responses to the same sentences, in order to test whether and where these processing stages occur in the human brain.

4. Phasic, ramping and sentence-final patterns in time-resolved multivariate decoding.

Applying multivariate decoding and temporal generalization to human sEEG (Figure 5A) and MEG recordings (Figure 5B) yielded a superposition of patterns. First, the sEEG and MEG diagonal decoding performance reached significance around 250 ms after word onset (Supplementary Figure 8; cluster permutation test on sEEG: significant cluster from 0.69 s to 4.10 s, $p=0.002$, and MEG: first significant cluster from 0.63s to 1.10s, $p=0.016$), consistent with studies comparing ERPs evoked by words and pseudowords (Poldrack et al., 1999; Woolnough et al., 2020). Decoding reached a peak around 500 ms after stimulus onset, reaching up to 0.65 AUC +/- 0.03 (SEM) for sEEG (Wilcoxon-Mann-Whitney test against chance: $p<0.01$) and 0.58 AUC +/- 0.03 for MEG (Wilcoxon-Mann-Whitney test against chance: $p<0.01$). The multivariate pattern that separated normal and Jabberwocky was similar across the four positions where content words appear in the sentence, resulting in a 4-by-4 grid of decoding generalization, similar to the one observed in neural language models (Figure 4). The resulting grid pattern means that even though the neural activity evolved over the course of the sentence, after each content word it transiently reached a similar state that dissociated normal and Jabberwocky sentences. This phasic pattern is consistent with a lexical process (Figure 1D).

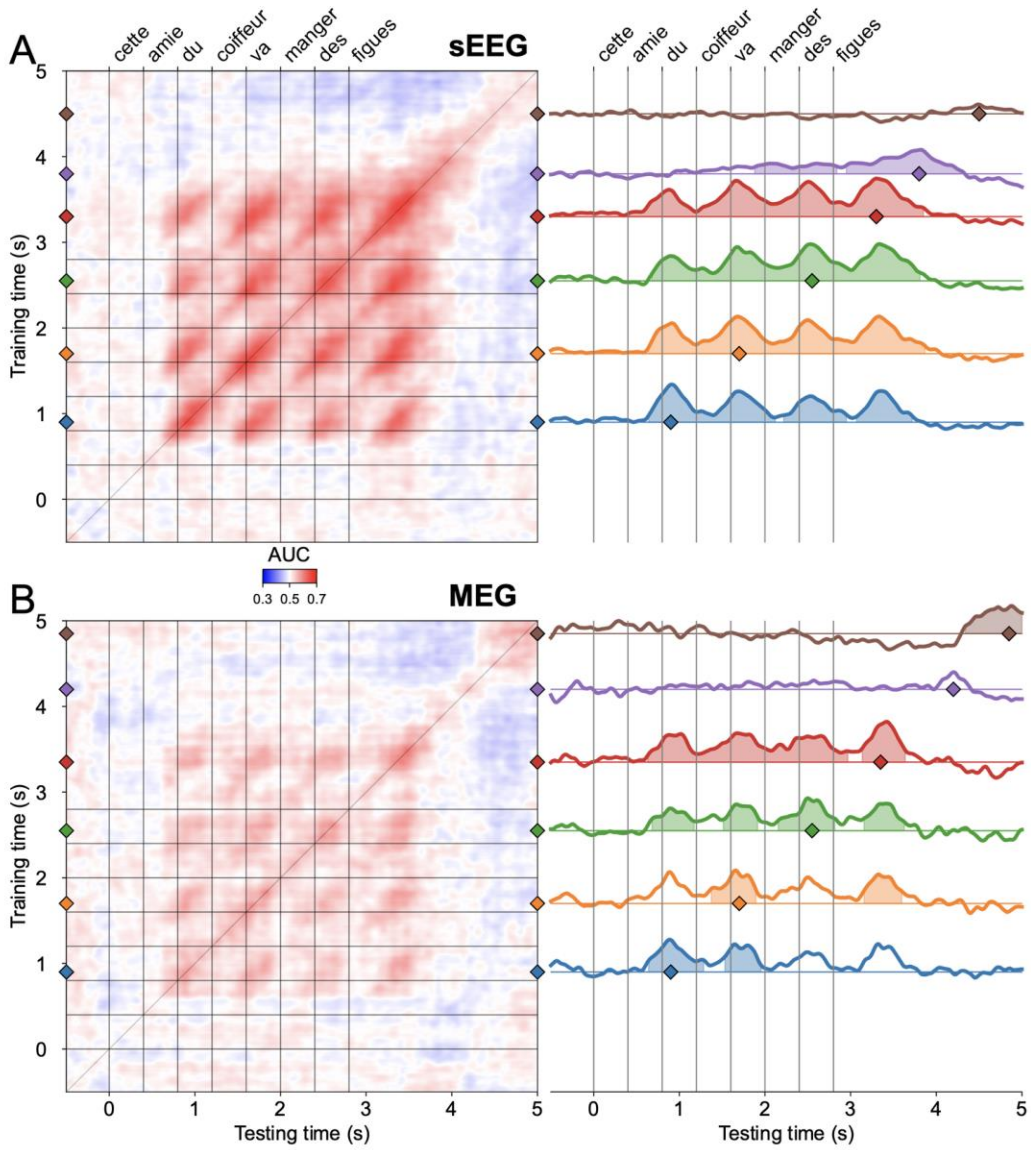


Figure 1-5: Decoding normal from Jabberwocky in human sEEG and MEG shows phasic, ramping and sentence-final patterns

Left: Temporal generalization matrices for a decoder trained to distinguish normal sentences from jabberwocky using in human sEEG (A) and MEG (B). The AUC is the average over the 11 subjects for sEEG and 9 subjects for the MEG.

Right: Generalization of individual decoders, i.e., horizontal slices from the temporal generalization matrices on the left. These slices from each matrix show in more details the temporal dynamics of sentence processing. Filled lines show significant time points, tested with cluster permutation test and FDR correction.

The first time point is the onset of the visual mask preceding the sentence. Each vertical grey line is a word onset. The small colored diamonds show the time where the decoders are trained.

Supplementary Figure 8 shows diagonal decoding performance and regression lines.

Second, to examine the presence of ramping effects, we tested whether decoding and generalization performance (average performance of a decoder over all timepoints) increased with sentence unfolding (from 0.4 s and 4 s). We found a positive effect in sEEG diagonal performance (linear regression on the average AUC across subjects: slope=0.013, r=0.39, p<0.0001), and the generalization performance (slope=0.0068, r=0.61, p<0.0001). In MEG, there was a significant effect for the diagonal performance (slope=0.0053, r=0.25, p<0.0001), but no effect for the generalization performance (slope=0.000072, r=0.014, p=0.81).

Finally, decoding performance remained significant for more than one second after the end of the sentence (cluster permutation test on sEEG: significant cluster from 0.69 s to 4.10 s, p<0.01, the last word's onset being at 2.8 s, Supplementary Figure 8). For example, the purple line in Figure 5A corresponds to a sEEG classifier trained 1 s after the last word's onset. Despite being trained this late, it reached an AUC of 0.59+/-0.03 and generalized to a few seconds before, with a clear ramping pattern (cluster permutation test: 2 significant clusters from 1.9 s to 2.8 s, p=0.014, and from 3 s to 4.3s, p<0.01). Similarly, the MEG classifier trained 2 s after the onset of the last word (t=4.9 s, Figure 5B brown line), hence much later than sensory and lexical processes, still showed a high decoding performance (training time AUC =0.56+/-0.03, cluster permutation test: one significant cluster from 4.3 to 5 s, p<0.01), now quite restricted over time and hence supporting the sentence-final wrap-up hypothesis (Figure 1D).

To summarize, at the whole brain level, we observed a linear superposition of the 3 patterns (Figure 1D) hypothesized to participate in semantic composition. Together, these results suggest that the brain integrates semantic information across multiple words and, after the end of the sentence, reaches a state that still differentiates normal and Jabberwocky sentences for a long period of time.

5. Superposition and regional specialization of phasic, ramping and sentence-final effects

To quantify the extent to which each of the three dynamic patterns was present in the empirical generalization matrices, we fit a linear regression using the (linearized) template matrices as predictors:

$$\hat{y} = \beta_{phasic} \times M_{phasic} + \beta_{ramping} \times M_{ramping} + \beta_{sentence-final} \times M_{sentence-final}$$

In this equation, $\hat{y}$ is the predicted AUC matrix, M_{phasic}, $M_{ramping}$ and $M_{sentence-final}$ are the template matrices shown in Figure 1D, and the betas are the corresponding estimated coefficients. To account for varying time delays and intrinsic dynamics, we performed a systematic grid search over template matrices, individually varying the onset and width of each peak (Supplementary Figure 9). The model with highest likelihood was selected. We thus obtained a beta coefficient for each template, quantifying the degree to which the dynamic pattern was present in the empirical matrix (averaged over all subjects). Note that this analysis is coarse: because we average patients' data before fitting the regression, only consistent patterns across patients will show up. The reason for this averaging is the small number of patients (11) and the fact that each patient does not have electrodes in every region, thus making the total number of data points too small for fitting a regression per patient followed by statistical testing across patients. Instead, we assess statistical significance with permutation tests (by shuffling the classifier's labels at the time of training the decoder).

Applying this method to the whole brain temporal generalization matrix (Figure 6B), we obtained significant coefficients (i.e., bigger than the coefficients fit on the AUC matrices from shuffled labels) for the three patterns: $\beta_{phasic} = 0.11$ ($p < 0.01$), $\beta_{ramping} = 0.06$ ($p < 0.01$), $\beta_{sentence-final} = 0.12$ ($p < 0.01$), confirming the findings of the previous section. The corresponding template matrices are shown in Figure 6A.

To evaluate whether the three dynamics revealed by temporal generalization arise from distinct brain regions, we repeated this regression analysis on subsets of electrodes belonging to anatomically defined regions of interest (ROIs), and predicted to be involved in distinct language-specific processes (Hickok & Poeppel, 2007; Friederici, 2011; Hagoort,

2019; Matchin & Hickok, 2020). We then plotted these results for each ROI by assigning a red, green and blue value corresponding respectively to the beta (normalized across regions to be between zero and one) of the phasic, ramping and sentence-final patterns. Example empirical matrices are shown in Figure 6C and the resulting whole brain maps in Figure 6D and 6E.

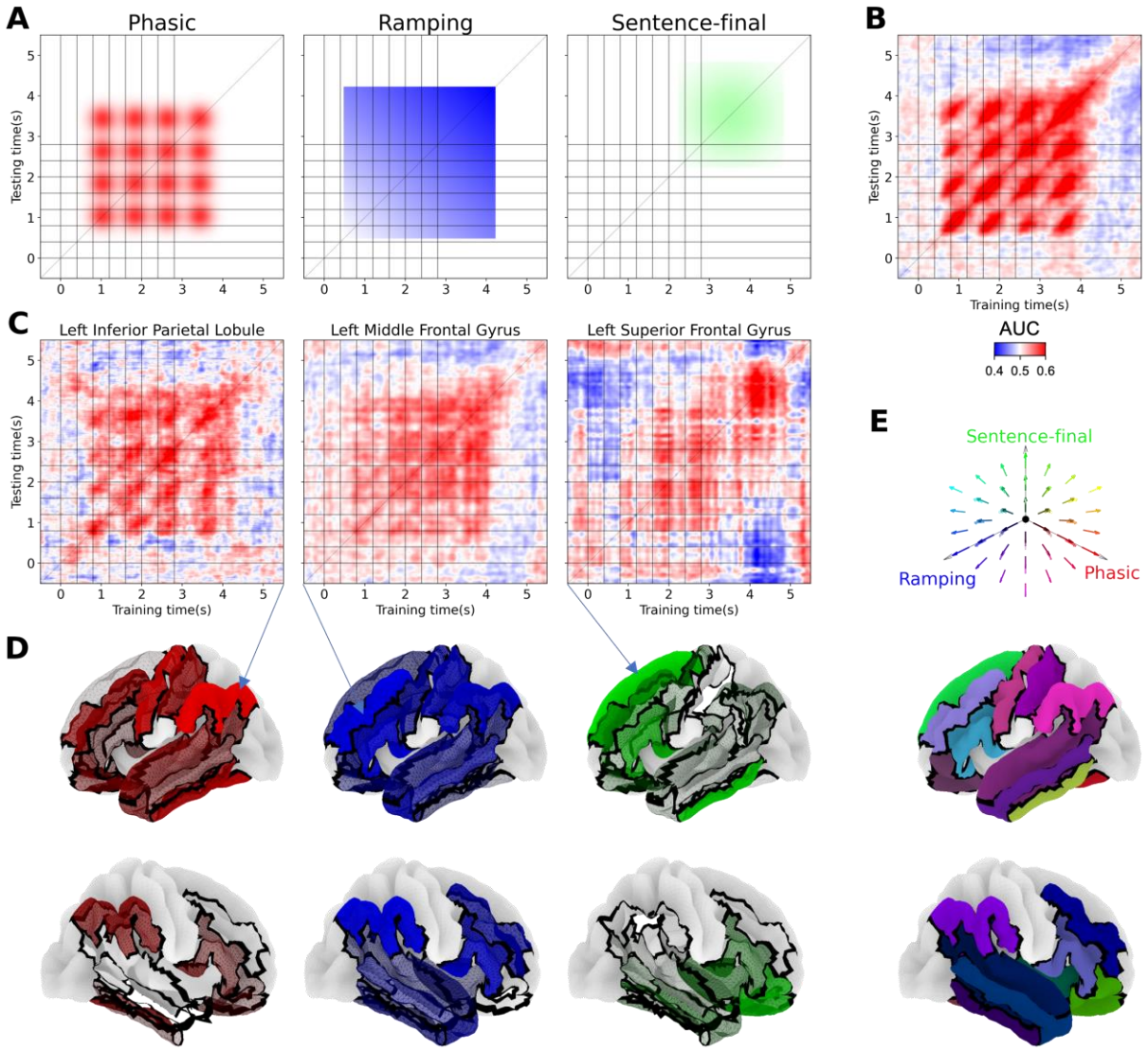


Figure 1-6: Phasic, ramping and sentence-final patterns are found in to varying degrees in each region in human sEEG

A. Template matrices selected by the grid search for whole-brain human sEEG. For each template, 20 different template matrices with varying delays and widths (see Supplementary Figure 9) were tested against the data, and the best fit was kept.

B. Empirical matrix for whole-brain human sEEG.
C. Empirical matrices in human sEEG for three most relevant ROIs. Arrows show the corresponding brain regions.
D. Surface brains maps showing the strength of the temporal generalization patterns for phasic (red, left), ramping (blue, middle), and sentence-final (green, right) processes in each ROI. The brain maps show the corresponding red, blue or green value, with a transparency value proportional to the regression coefficient of the corresponding pattern.
E. Surface brain map showing the combined strength of the 3 patterns in each ROI. The color of each ROI reflects the red, blue, and green values of the phasic, ramping and sentence-final patterns.

Supplementary Figure 9 shows the range of templates used in the gridsearch regression.
Supplementary Figure 10 shows the lack of syntactic modulation in whole brain decoding performance.

First, we expected the ventral occipito-temporal visual pathway, and particularly the FuG, to exhibit phasic (lexical) effects, with no ramping or delayed patterns. Our results are broadly consistent with this prediction as shown by the strong presence of the lexical component ($\beta_{phasic} = 0.84, \mathrm{p} < 0.01$), but not ramping ($\beta_{ramping} = 0.14, \mathrm{p} = 0.12$) nor sentence-final ($\beta_{sentence-final} = 0.07, \mathrm{p} = 0.09$) components in the FuG (Figure 6D and 6E). To verify that the ramping pattern was not present in these regions, we checked that the performance increased only marginally and non-significantly over the course of the sentence (linear regression AUC for each subject: average slope=0.0018, Wilcoxon Mann-Whitney test against null slope: p=0.43 for the left precentral gyrus).

Second, we expected the MTG and STG, to be responsible for accessing lexical representations (Hart et al., 2000; Binder et al., 2003; Tranel, 2009) and starting to compose them according to sentential context (Lau et al., 2008; Pallier et al., 2011; A. R. Price et al., 2015, 2016), thus showing a combination of the phasic and ramping patterns. Our results are consistent with this prediction (left STG: $\beta_{phasic} = 0.54, \mathrm{p} < 0.01$, $\beta_{ramping} = 0.53, \mathrm{p} < 0.01$, $\beta_{sentence-final} = 0.17, \mathrm{p} = 0.1$; left MTG: $\beta_{phasic} = 0.42, \mathrm{p} < 0.01$, $\beta_{ramping} = 0.66, \mathrm{p} < 0.01$, $\beta_{sentence-final} = 0.12, \mathrm{p} = 0.08$): the normal versus jabberwocky decoding results revealed a strong response starting 200 ms and peaking 400 ms after each content word onset (e.g. peak performance for the first word at t=0.8 s: mean AUC=0.56 +/- 0.02 Wilcoxon-Mann-Whitney test against chance: p<0.01), with good generalization to all words in the sentence along with an increase of performance over time (linear regression for each subject: average slope=0.0039, Wilcoxon Mann-Whitney test against null slope: p<0.01) and an increase of the generalization performance (average

slope=0.0047, Wilcoxon Mann-Whitney test against null slope: p<0.01). Interestingly, parietal regions such as the IPL showed a similar pattern (figure 6C left): $\beta_{phasic} = 1, \mathrm{p} < 0.01$, $\beta_{ramping} = 0.83, \mathrm{p} < 0.01$, $\beta_{sentence-final} = 0.18, \mathrm{p} = 0.07$, confirming their involvement in word composition (Bemis & Pylkkänen, 2013a; A. R. Price et al., 2015, 2016).

Third, we expected the prefrontal cortex to be more specifically involved in combinatorial computations (Hagoort, 2005; Pallier et al., 2011; Friederici, 2011; Fedorenko et al., 2016; Nelson et al., 2017; Matchin & Hickok, 2020). In the left IFG ($\beta_{phasic} = 0.2, \mathrm{p} = 0.08$, $\beta_{ramping} = 0.85, \mathrm{p} < 0.01$, $\beta_{sentence-final} = 0.69, \mathrm{p} < 0.01$) and in the left MFG (Figure 6C middle; $\beta_{phasic} = 0.64, \mathrm{p} < 0.01$; $\beta_{ramping} = 1, \mathrm{p} < 0.01$, $\beta_{sentence-final} = 0.58, \mathrm{p} < 0.01$), we observed a ramping activity profile, starting around 350 ms after the first content word's onset, and increasing without discontinuity until 4.2s, i.e., 1.4 s after the last word onset (linear regression for each subject: average slope=0.0065, Wilcoxon Mann-Whitney test against null slope: p<0.001 for left IFG). The ramping pattern was also visible in the generalization performance (average slope=0.0054, Wilcoxon Mann-Whitney test against null slope: p<0.01). Similar ramping profiles were found in the right IFG and MFG (Figure 6D and 6E), but the evidence for phasic and sentence-final patterns was weaker. This activity profile is consistent with a linear integrator (Pallier et al., 2011; Fedorenko et al., 2016), whereby I/MFG would combine each incoming word with the previous ones.

Last, in the right OFC, ($\beta_{phasic} = 0.26$, p=0.15, $\beta_{ramping} = 0, \mathrm{p} = 0.96$, $\beta_{sentence-final} = 0.71, \mathrm{p} < 0.01$), and, the left SFG (Figure 6C, $\beta_{phasic} = 0.09, \mathrm{p} = 0.47$, $\beta_{ramping} = 0.34, \mathrm{p} = 0.05$, $\beta_{sentence-final} = 1, \mathrm{p} < 0.01$), decoding performance stayed at chance level for the most part of the sentence, but significantly increased after the last word, and stayed above chance until 1.6 s after the end of the sentence (cluster permutation test on the diagonal performance: single significant cluster from 3.9 s to 4.4 s, p<0.01 for left SFG and single significant cluster from 4.0 s to 4.4 s, p<0.01 for right OFC). This sentence-final effect is consistent with a wrap-up process.

In sum, we successfully identified a set of regions exhibiting signatures of phasic, ramping and sentence-final processes, and thus provide a path to a systematic decomposition of sentence composition in the brain.

E. Discussion

To clarify how sentences are composed by the human brain, we introduced and tested a simple yet powerful vector coding framework in NLMs and human electrophysiological recordings. First, based on general arguments on the use of increasingly large neural subspaces to encode the compositional meaning of sentences, we predicted that the dimensionality of brain signals should increase as successive words get added to an evolving representation of sentential meaning, and that this effect should be larger for meaningful than for meaningless materials (Jabberwocky or lists of meaningless strings). In agreement with this prediction, we found that representations of meaningful sentences evoke neural signals of higher dimensionality than Jabberwocky sentences or string lists. It is noteworthy that there was an increase of dimensionality with time in all conditions - including string list – but, crucially, it was highest for normal sentences. Surprisingly, the increase in intrinsic dimensionality started relatively late in the MEG, ~1 s after the increase started in sEEG. This finding may suggest that the effect starts locally before being propagated to the whole brain. These effects were absent in untrained NLMs, supporting the idea that, with learning, coding dimensions are assigned distinctive meaning in distributed semantic spaces.

To further characterize the dynamics of brain activity during sentence processing we used multivariate decoding and temporal generalization, a method that has recently been advocated for in the context of language processing (Fyshe, 2020; He et al., 2022). The results indicate that the representations generated in brains and NLMs follow similar dynamics, despite their differences in implementation and timescale. Specifically, we observed three distinct dynamic signatures: phasic, ramping and sentence-final, which we interpret as the reflection of single-word processing, multi-word composition, and sentence wrap-up, respectively. ROI analysis showed that some regions, such as the FuG, exhibited pure lexical patterns, thus confirming its role in written word identification. On the other hand, regions such as MTG, STG and IPL, displayed mixed lexical and ramping patterns, suggesting that they come first in the compositional process, having dynamics compatible with both lexical access and multi-word integration. Frontal regions had mixed signatures of ramping and sentence-final (for IFG), as well as a phasic component (for MFG), or a pure

sentence-final effect (for right OFC and left SFG). Overall, we observe a wide variety of combinations of each signature, witnessing the dynamic flow of information during sentence processing. Although this pattern-based analysis is coarse and is correlational rather than causal, it is corroborated by single-channel evoked activities (e.g. Figure 2) and fits with the previous literature.

Lexical access, in particular, has been studied extensively and is thought to be supported by the FuG and temporal regions , where we found strong phasic signatures. Curiously, this was also the case of parietal regions such as IPL and the pre and postcentral gyri, suggesting a stronger involvement in lexical access than previously thought. The ramping pattern also appeared as a marker of compositional processes in several previous studies. Pallier and colleagues (2011) observed that fMRI activity in IFG and posterior Superior Temporal Sulcus (STS) increased in direct proportion to the number of elements in the current syntactic phrase, for both normal text and Jabberwocky. They proposed a simple model in which each consecutive word or phrase adds a fixed amount of activity to a compositional representation which therefore builds up across time. Nelson and colleagues (2017) and Fedorenko et al. (2016) then showed, with the higher resolution of intracranial EEG, that high-gamma activity does indeed increase after each word in a constituent word phrase.

The exact computational role of this ramping activity is, however, still unknown. Since we studied the contrast between normal and Jabberwocky sentences, which primarily differ in semantic but not syntactic content, we interpret it as a marker of the combination of each incoming word into the growing combinatorial representation of sentential meaning. Due to these semantic processes, the neural assemblies recruited during the processing of normal sentences should be larger and the neural dynamics richer, compared to Jabberwocky. Computational models of the neural encoding of compositional structures (Smolensky, 1990; Plate, 1995; Gayler, 2004) indeed predict that the neural code can be characterized as a sum of representations for each constituent (technically, a sum of tensor products of the vectors representing each word's role and filler) and should therefore increase as their number increases. Nevertheless, we acknowledge that direct evidence for such a neural code is still missing. Testing whether brain activity during sentence processing contains signatures of tensor product representations is a promising avenue for future work.

Interestingly, previous studies found that RNNs trained on artificial grammar tasks learn representations compatible with the tensor product framework (McCoy et al., 2018; Soulos et al., 2020).

The sentence-final pattern seen in SFG and OFC may be associated with several cognitive processes. The classical view of wrap-up processes is that they reflect higher-level integration and if necessary, reanalysis and conflict resolution of the multiple possible meanings of words (Just & Carpenter, 1980; Molinaro et al., 2008). It may also indicate delayed composition, whereby the ultimate meaning of the sentence would only be composed once all words have been read; or a memory process that stores this meaning and holds it in working memory. Our design does not allow to distinguish those interpretations. Until recently, few neuroimaging studies examined sentence-final activations, for fear that wrap-up effects would confound regular processes happening at the last word (Stowe et al., 2018). By contrast, in our results, signatures of wrap-up are found in distinctive brain regions, thus suggesting that they can be studied independently. We hope that this will encourage more studies of the computations underlying wrap-up effects.

Note that, here, we also used varied syntactic structures (Figure 1C) as an initial attempt to look for modulations of the dynamic patterns by syntax. However, no such modulations were found (Supplementary Figure 10), suggesting that the integrative processes are similar in the three structures used. This does not preclude a modulation in more complex structures, for example sentences including adverbial phrases or embedded clauses.

The present intrinsic dimensionality hypothesis stems from much research in the past decades, which has emphasized how biological neural network use high-dimensional vector spaces to encode complex structures (Quiroga et al., 2005; Tyukin et al., 2019; Gorban et al., 2019; Calvo Tapia et al., 2020). For instance, high-dimensional vectors are used in some tensor product theories of neural composition (Smolensky, 1990; Plate, 1995; Eliasmith & Anderson, 2003; Smolensky et al., 2022) where distributed neural vectors representing each of the sentence's constituents and their thematic roles are summed, leading to progressive divergence from the null vector, thus resulting in an increase in

intrinsic dimensionality. A related idea is that the brain uses different parts, or "subspaces", of the huge vector space made available by the independent activity of millions of neurons to store the different elements of an incoming sequence and their relationships. This assignment of dimensions has been demonstrated directly both in artificial neural networks (Advani et al., 2020; Lakretz et al., 2019) and at the brain level (Liu et al., 2019; Flesch et al., 2022; Xie et al., 2022). For instance, human MEG signals reflect the multiple, factorized dimensions of an ongoing visual sequence (Liu et al., 2019; Quentin et al., 2019; Al Roumi et al., 2021), and recordings of thousands of monkey prefrontal neurons can be decomposed into orthogonal vector subspaces storing the successive elements of a spatial sequence in working memory (Xie et al., 2022). Such findings could be extended using the present intrinsic dimensionality analysis as well as non-linear alternatives (Facco et al., 2017; Granata & Carnevale, 2016; Landa et al., 2021). The advantages of the measure of intrinsic dimensionality used here are its simplicity and wide acceptance. This notion of intrinsic dimensionality has been used for some time in statistics and machine learning (Campadelli et al., 2015; Carreira-Perpinán, 1997) and has recently gained traction in the neuroscience domain, where low-dimensional intrinsic dynamics were found in high-dimensional neural recordings (Machens et al., 2010; M. M. Churchland et al., 2012; Mante et al., 2013; Xie et al., 2022). These low-dimensional dynamics have also been found to emerge in trained neural networks (Laje & Buonomano, 2013; Recanatesi et al., 2021) and have been assigned important roles in various theories of neural computation (Gallego et al., 2017; Gao et al., 2017; Vyas et al., 2020; Ebitz & Hayden, 2021; Sorscher et al., 2021).

Regarding artificial neural networks, an influential study by Mikolov and colleagues (Mikolov, Sutskever, et al., 2013) showed that single words can be represented by dense high-dimensional vectors in semantic spaces learned from word co-occurrence statistics in large corpuses. The dimensions of such semantic spaces map onto interpretable semantic features such as gender, location, and size (Senel et al., 2018). These distributed word representations were later shown to align with cortical responses to words by means of linear encoding models (Grand et al., 2022; Huth, Lee, et al., 2016). Similar work has extended these results, showing that the cortical responses to natural language can be mapped onto the vector representations extracted from NLMs when presented with the same sentences (Jain & Huth, 2018; Toneva & Wehbe, 2019; Caucheteux & King, 2020;

Caucheteux et al., 2021a; Schrimpf et al., 2021; Goldstein, Zada, et al., 2022). These similarities, together with the fact that NLMs perform somewhat similarly to humans on various linguistic tasks (Otter et al., 2021), and make similar errors (Coenen et al., 2019; Goldberg, 2019; Jawahar et al., 2019; Lakretz et al., 2020; Lakretz, Desbordes, King, et al., 2021a), provide an interesting, although incomplete, means to study neural representations of sentences during language processing in a system which is (1) less noisy compared to brain data, (2) fully accessible to manipulation and recordings, and (3) not a priori tied to a normative linguistic theory but rather, learned from large-scale linguistic corpora. For the first time, we present a complementary decoding approach to show that the similarities between brains and artificial neural networks lie not only in their representations, but also in their dynamics (i.e., the order in which the representations are combined with one another). Perhaps even more surprisingly, the resemblance holds for non-canonical stimuli such as Jabberwocky sentences, which are out-of-domain for both brains and NLMs. The convergence is intriguing, especially given that the character-based NLMs used in this study are still far from human-level language understanding and are trained on very large corpora to predict characters from surrounding contexts – a learning scheme at odds with language acquisition in humans. Of note, the interpretation of the sentence-final pattern in the models is not straightforward and we remain careful to interpret it as a wrap-up process. Contrary to humans, the models are trained to predict an output at each time point, and thus it may be improbable that the models keep individual words in memory in order to combine them later. The only thing we can say for sure is that the models keep in memory the compositional meaning of the preceding sentence(s), and that this is reflected in their internal representations.

A limiting factor in our work is the limited coverage of some brain regions (e.g., STS). Furthermore, we had to pool together electrodes in relatively large brain regions in order to achieve decent decoding performance. This choice greatly limited the spatial accuracy of our analyses. For example, subregions of the IFG are known to have functional specializations and inter-individual variability (Fedorenko & Blank, 2020): the pars triangularis (BA45) was found to be most sensitive to syntax (Nelson et al., 2017), whereas the pars opercularis (BA44) is also involved in tasks involving monitoring and sequencing speech sounds at the phonological level (Zatorre et al., 1992; Poldrack et al., 1999). Similarly, we had to aggregate

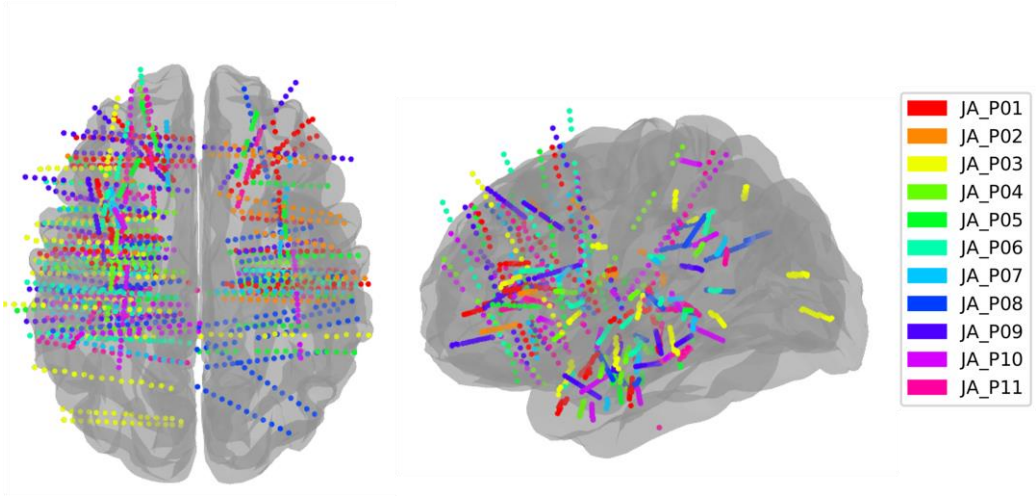
the temporal pole, which has been associated with two-word composition (Bemis & Pylkkänen, 2011; Fyshe et al., 2019; Pylkkänen, 2019), with the temporal gyri; and our analyses also conflated the angular gyrus, which has been associated with semantic composition (A. R. Price et al., 2015, 2016) within a vast region that we labeled as inferior parietal lobule (IPL). Higher-resolution recordings, for instance using smaller electrode arrays (Steinmetz et al., 2021; Szostak et al., 2017), will be needed in order to further study the functional specialization of these regions. Furthermore, our study considers semantic composition in the broadest sense and does not afford any claim regarding the specific subprocesses underlying the observed dynamics. Specifically, our design cannot separate logico-semantic from conceptual combinations (Pylkkänen, 2019), which are confounded in our main contrast of normal versus Jabberwocky sentences.

Taken together, our results suggest that a succession of processing stages, separated by their distinct brain signatures, underlie the composition of sentence-level semantics. They allow us to speculate that incoming lexical information arising from the FuG (Woolnough et al., 2020) is first passed to the temporal lobe and inferior parietal lobule, where semantic information is accessed, stored, and begins to be combined. The I/MFG exhibits relatively selective ramping and sentence-final signals (Figure 6D), suggesting that it may play a key role in merging individual words into constituents that encode their compositional meaning. The final sentential representation may then stay present in the activity pattern of the left SFG and right OFC for several seconds – a duration which might have been extended if we had presented multiple sentences forming part of a larger discourse. Finally, the construction of these compositional meanings is associated with an increased dimensionality of the representations. These results bring us one step closer to understanding how the human brain composes and understands sentences.

F. Acknowledgements

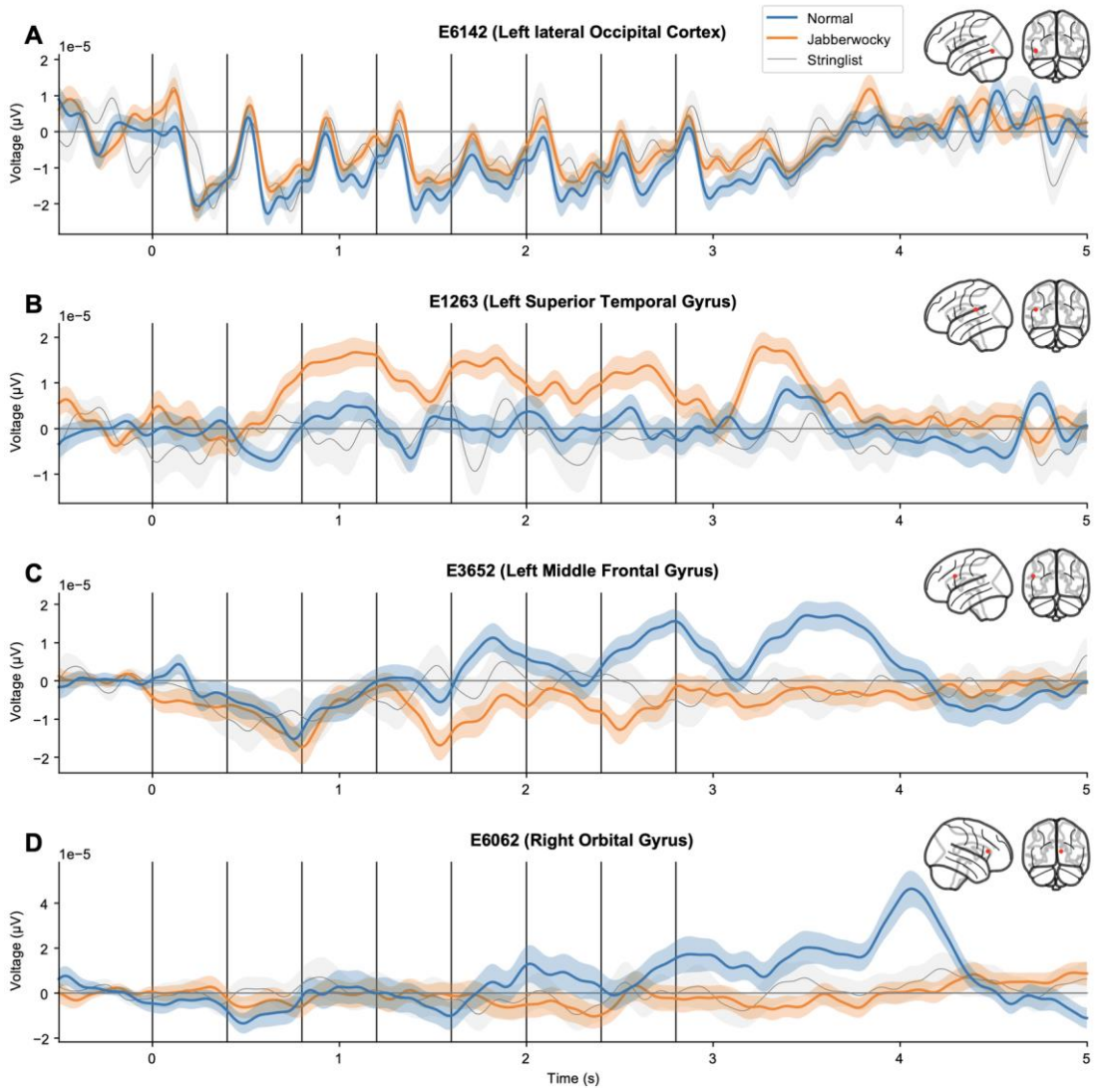
Researchers from the Institute of Language, Communication and the Brain (ILCB) were supported by grants ANR-16-CONV-0002. Christian Bénar was partly funded by a FLAG ERA/HBP grant from Agence Nationale de la recherche "SCALES" ANR-17-HBPR-0005. Stanislas Dehaene was funded by INSERM, CEA, Collège de France, the Bettencourt-Schueller foundation and an ERC grant “NeuroSyntax”. Data acquisition was performed on a platform member of France Life Imaging network (grant ANR-11-INBS-0006).

G. Supplementary Figures



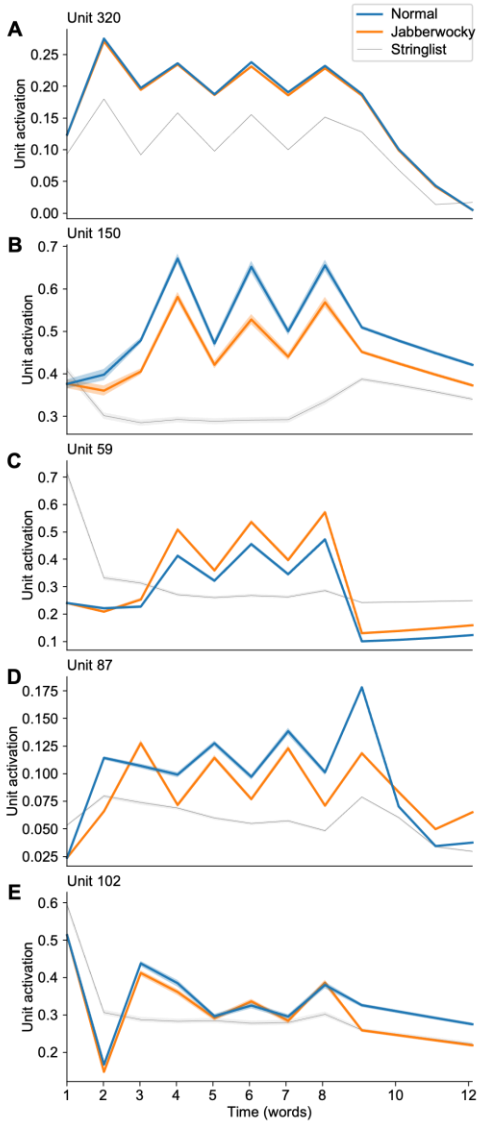
Supplementary Figure 1: Intracranial electrode coverage

Electrode location in Montreal Neurological Institute (MNI) space. Each patient is shown with a different color.



Supplementary Figure 2: Additional illustrative profiles of human sEEG responses

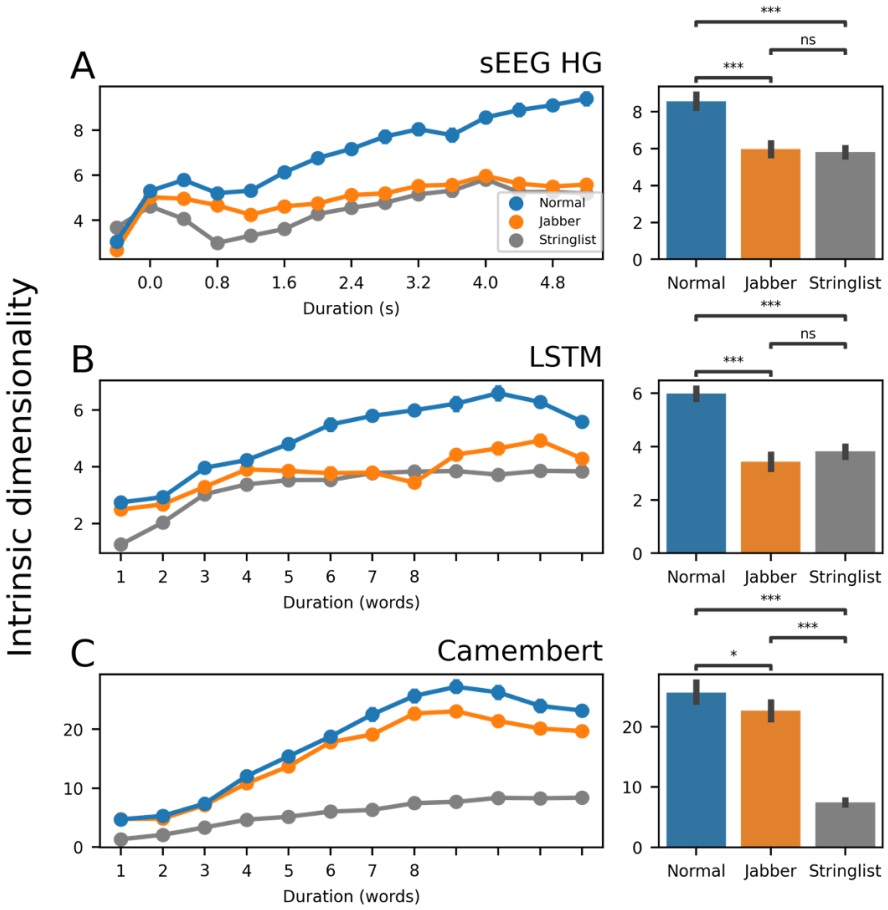
Four examples of electrodes responding to normal sentences (blue), Jabberwocky sentences (orange), and string of consonants (grey). Each line shows the local field potential (Voltage) for each electrode.



Supplementary Figure 3: Diversity of dynamics in units from the Transformer's last layer

Average activation from hand-picked units in the Transformer's 12th layer, showing interesting dynamics, sometimes surprisingly similar to human sEEG activations.

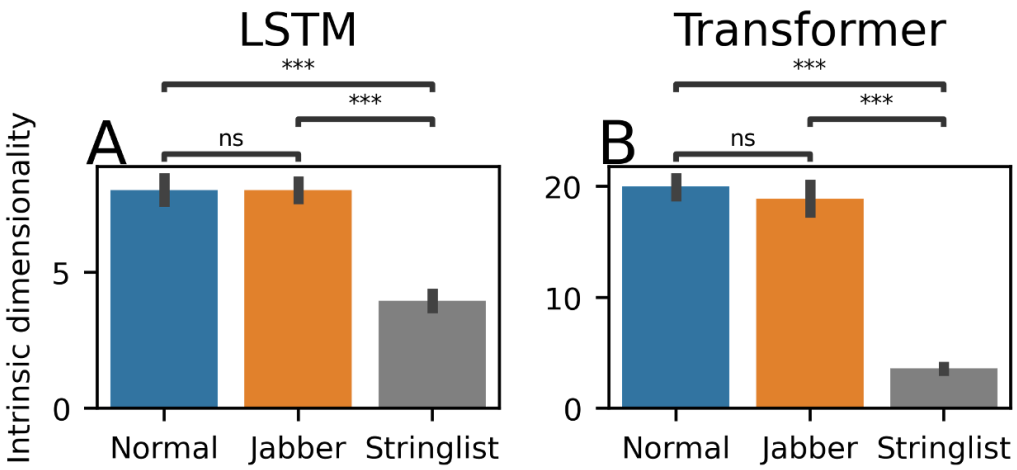
A. Orthographic effect. Word length affects the activations (higher for longer words).
B. Lexical effect. Activation differs between normal and Jabberwocky after the first content word.
C. Same as B but with higher activations in the Jabberwocky condition.
D. Ramping effect. The difference between normal and Jabberwocky increases over the course of the sentence.
E. Sentence-final effect. Although the activations are similar during the sentence, normal and Jabberwocky diverge after the presentation of the last (8th) word.



Supplementary Figure 4: Intrinsic dimensionality is higher for normal sentences than Jabberwocky

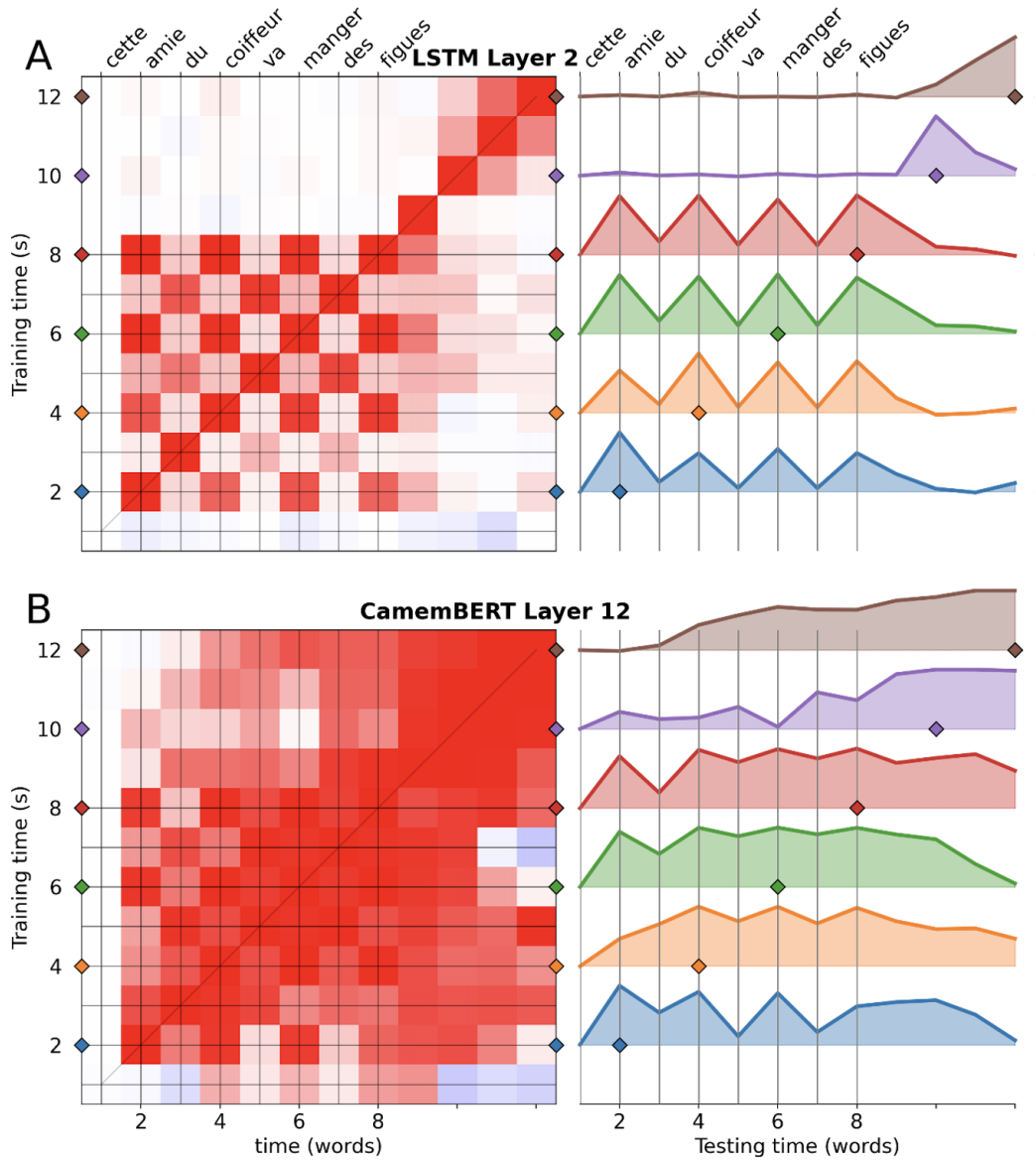
A: Intrinsic dimensionality of the sEEG high gamma signals from all subjects, as a function of the time window used (sliding time window of 0.4 s width).
Right: Bar plot showing the intrinsic dimensionality computed using the whole sentence (4 s time window).

B, C: same analysis applied to LSTM and CamemBERT activations (last layer, averaged over all stimuli for each condition). The bar plots show the intrinsic dimensionality computed using the 8 words time window.



Supplementary Figure 5: Untrained language models do not exhibit a larger intrinsic dimensionality for normal versus Jabberwocky sentences

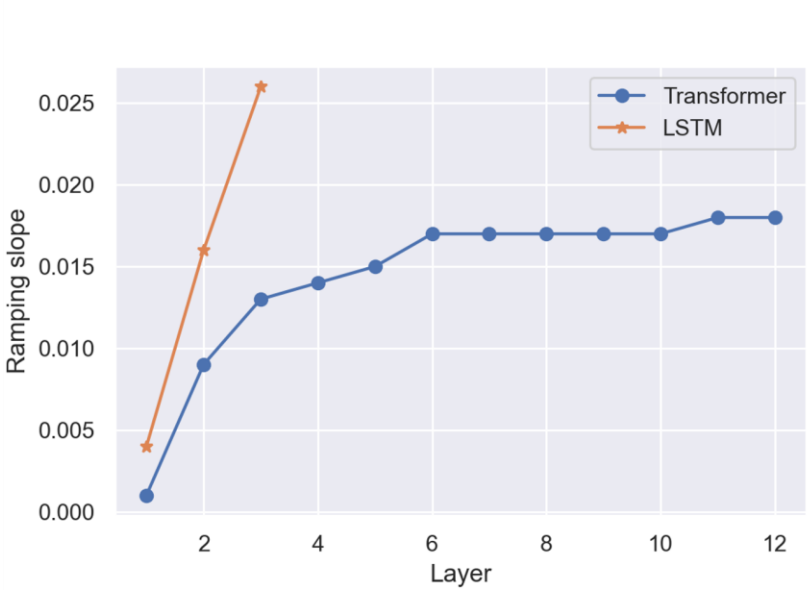
The figure shows the Intrinsic dimensionality computed using the whole sentence (8 words) time window for an untrained LSTM (A) and an untrained Transformer (B). There was no significant difference between the normal and Jabberwocky conditions.



Supplementary Figure 6: Decoding normal versus Jabberwocky sentences in LSTMs and CamemBERT

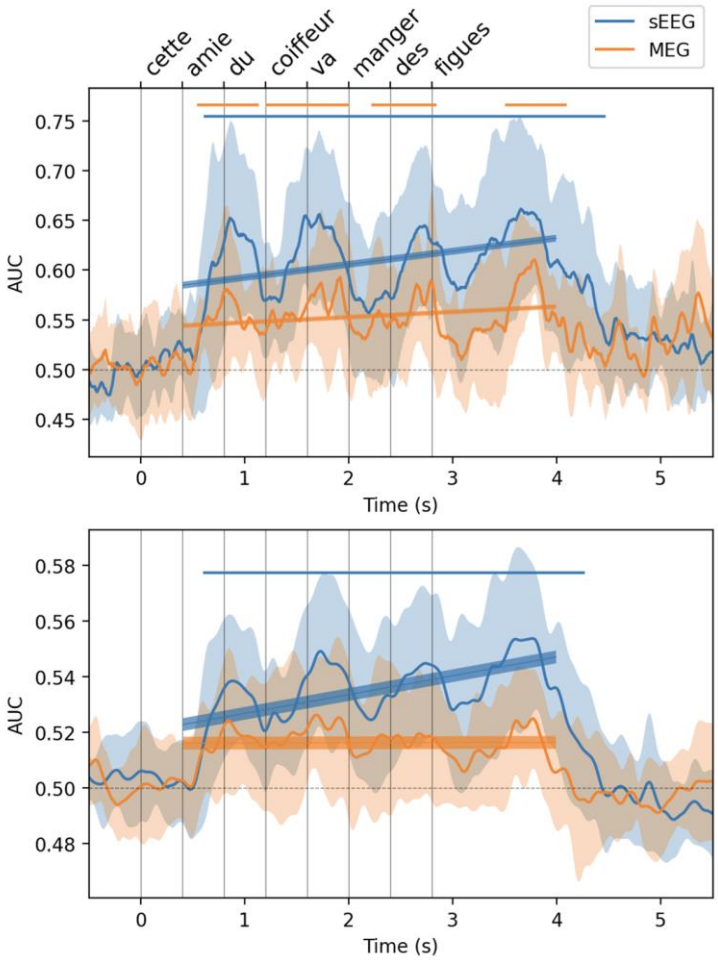
Left: Temporal generalization matrices for a decoder trained to distinguish normal sentences from jabberwocky using the activity of an LSTM's 2nd layer (top) and the CamemBERT's 12th layer (bottom).

Right: Generalization of individual decoders, i.e., horizontal slices from the temporal generalization matrices on the left. These slices from each matrix show in more details the temporal dynamics of sentence processing. Filled lines show significant time points, tested with Wilcoxon-Mann-Whitney test against chance (0.5) and FDR correction.



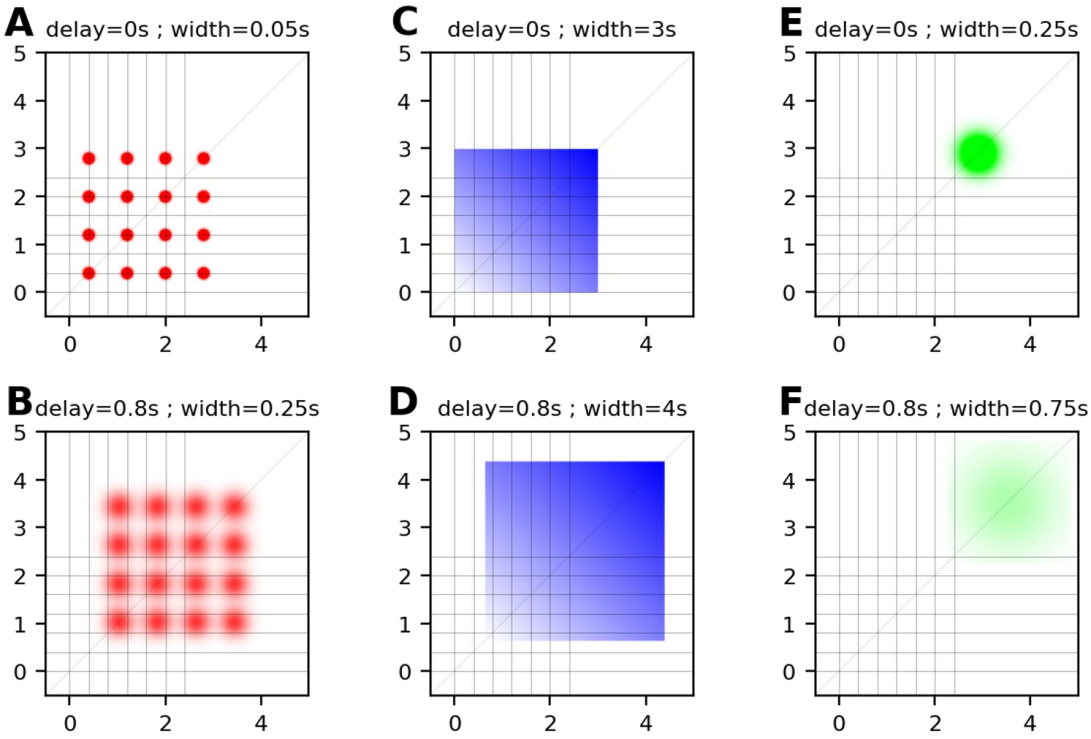
Supplementary Figure 7: Ramping tendency in LSTMs and Transformers NLMs

Slope of the linear regression on the classification performance between word 1 and 8, for each layer of the LSTMs and Transformers NLMs under study. The ramping slope increases with the layer number. We included all 10 instances of each model in this regression to get substantial statistical power. Because of this, we could not apply this analysis to the CamemBERT model for which a single instance was available.



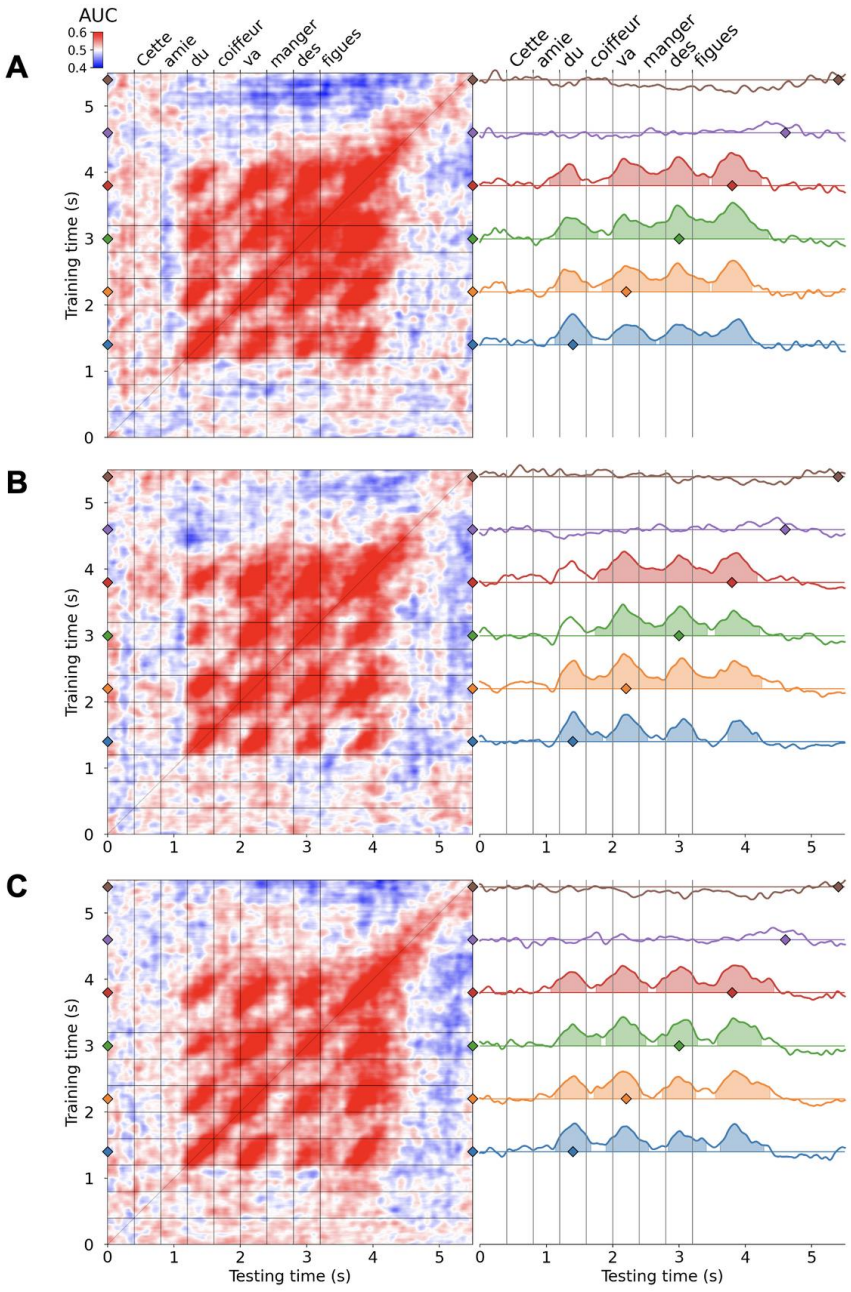
Supplementary Figure 8: Diagonal decoding performance for normal versus Jabberwocky sentences in human sEEG and MEG

Within-time (diagonal) decoding performance (A) and generalization performance (B), i.e., the average of each line of the temporal generalization matrix. Filled lines show significant time points, tested with cluster permutation test and FDR correction. Regression lines and 95%confidence intervals are also shown. Note that performance stays above chance for a long time after the last word was presented.



Supplementary Figure 9: Example templates used in the grid search for the template regression analysis

Matrices shown are the extremes for each parameter (delay and width) selected by the grid search. Each parameter was varied independently of the other. For each template, we tested 10 delays and 10 widths, covering the most likely range of dynamics of interest.



Supplementary Figure 10: Lack of syntactic modulation of the ramping pattern

The figure shows the temporal generalization matrices of decoders trained on normal versus Jabberwocky using all syntactic structures, and then tested on each structure separately. Matrices correspond to syntactic structure 2-6 (A), 4-4 (B) and 6-2 (C). We tested whether the transition from NP to VP would induce a decrease in decoding performance (peak performance just before the transition, i.e., at the last noun of the NP versus just after the transition, i.e., at the verb, using Wilcoxon-Mann-Whitney tests, but none of these effects were significant.

Introduction to chapter 2

Following this first study, for which the data was already available, we strove to make a new experimental design that could separate the representation of words when they are presented in isolation and when they have to be combined with other words. As described in the introduction, many studies dissected the neural mechanisms underlying two-word compositions (Pylkkänen, 2019, 2020a). We sought to go further and included phrases of up to five words. After piloting this new paradigm, I recorded thirty participants with magnetoencephalography. I then implemented a combination of time resolved decoding, event related fields, and behavioral analyses to characterize the dynamics and format of the representations of phrases of increasing length.

Thus, for this study I participated in all the steps of academic research: inventing the experimental design, piloting it, recording the data, analyzing it, and writing the article that reports our finding.

I want to thank the pilots and participants of this study that endured multiple repetitions of each possible combination of the eight words used as stimuli.

An updated version of this manuscript has been submitted to the journal *Neuron* and is still in review at the time of writing.

Chapter 2. Evidence for a compressed neural code in working memory during language composition

Théo Desbordes*, Meta AI Research, Paris, France & Cognitive Neuroimaging Unit NeuroSpin center 91191, Gif-sur-Yvette, France

Jean-Rémi King, PSL University, CNRS & Meta AI Research, Paris, France

Stanislas Dehaene, Université Paris Saclay, INSERM, CEA, Cognitive Neuroimaging Unit, NeuroSpin center, Saclay, France ; and Collège de France, PSL University, Paris, France

A. Abstract

The ability to compose successive words into a meaningful phrase is a characteristic feature of human cognition, yet its representational format and neural bases remain largely unknown. Here, we put forward several key mechanisms of compositionality using magnetoencephalographic (MEG) recordings of brain activity while participants compared 1-, 2- and 5-word phrases to a subsequent image. The decoding of MEG signals reveals three main findings: First, the representation of each word is partially sustained until it can be integrated into a coherent phrase, at which point it fades away. Second, the neural activity during the delay period increases with phrase complexity. Third, the speed and accuracy with which a phrase can be matched with a picture depends on phrase complexity and is faster for surface properties of the phrase compared to syntactically deeper ones. We suggest that compositional representations are compressed in working memory and require a period of decompression to be accessed and used for picture matching. Overall, these results shed new light on the nature of compositional representations in the human brain.

B. Introduction

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

There is however, no consensus regarding how the brain combines the meaning of individual elements, and how it represents such composition (Friederici et al., 2017; A. E. Martin & Doumas, 2017; Frankland & Greene, 2020a). Among these theories, vector-symbolic architecture such as the tensor-product representation (Smolensky, 1990) and the semantic pointer architecture (Eliasmith & Anderson, 2003; Eliasmith et al., 2012) have seen notable success. For example, the semantic pointer architecture has been put to use in theories of concepts (Blouw et al., 2016), emotions (Kajić et al., 2019) and consciousness (Thagard & Stewart, 2014). It was also found that representations of artificial neural networks could be well approximated by tensor-product representations when they are trained on artificial, explicitly compositional sequence-to-sequence tasks, but not when they are trained on natural language (McCoy et al., 2019; Soulos et al., 2020). Relatedly, recent successes in deep learning models of natural language processing seem to be partly due to their good generalization properties, although they do not rely on systematic compositional rules (Baroni, 2020; Brown et al., 2020b; Chaabouni et al., 2020). For example, even state-of-the-art image generation models from natural language, such as OpenAI's Dall-E 2 (Radford et al., 2021; Ramesh et al., 2022) and Google's Imagen (Saharia et al., 2022) can dramatically fails on simple compositional and binding operations (Conwell & Ullman, 2022; Marcus et al., 2022).

Multiple brain regions have been associated with compositional processes. Famously, "Broca's area", more precisely the pars opercularis and triangularis of the inferior frontal gyrus (IFG) has long been thought to be the siege of unification operations (Hagoort, 2005), including composition and binding. This region was repeatedly found to be more active in conditions where composition could happened, such as sentences versus word lists (Mazoyer et al., 1993; Humphries et al., 2005; Friederici et al., 2010), normal sentences

versus Jabberwocky (Fedorenko et al., 2016), and constituents of increasing size (Pallier et al., 2011; Nelson et al., 2017). In many of these studies, the posterior superior temporal sulcus was also found to be active.

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

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

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

This view is somewhat opposite to classical working memory theories where each characteristic of a stimulus is represented explicitly by sustained firing of specific neurons (Goldman-Rakic, 1995; Leung et al., 2002). This same goes for more recent activity-silent theories of working memory, which posit that change in network-level characteristics (such as short-term plasticity) are the basis for the storage of short-term memoranda (Mongillo et al., 2008; Stokes, 2015), but no constraint is set on the representational format of the

storage. Even recent proposal that combine sustained and activity-silent working memory (Spaak et al., 2017; Trübutschek et al., 2019; Barbosa et al., 2020; Stokes et al., 2020) consider that information is stored as is, not compressed, as suggested by Restle and others.

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

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

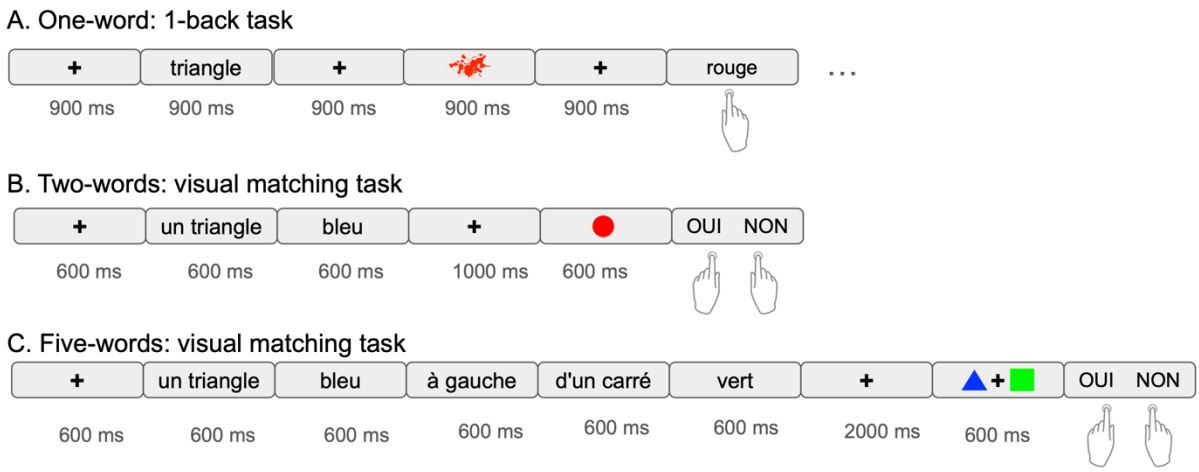


Figure 2-1: Experimental designs: hybrid 1-back and delayed sentence-to-image matching tasks

A: The one-word blocks consist of a 1-back across words and images. Subjects have to indicate, in a long series of random stimuli, when two stimuli represent the same meaning.

B, C: Delayed sentence-to-image matching task. In two-word blocks (B) and five-word blocks (C), subjects have to determine whether the object described by the two successive words matches the image presented 1 s later.

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

We consider three pairs of hypotheses:

- Hypothesis 1 : if an active representation of each element is necessary for online composition, then individual words should be explicitly represented until they are combined with their corresponding phrase. On the other hand, if composition can occur solely with activity-silent mechanisms, then the active representation of each word until the end of its constituent would not be necessary.
- Hypothesis 2: if neural representations are compressed for short-term storage, then some neural signal should reflect the complexity (i.e., the quantity of information) of the sentence. To the contrary, if the representations are not compressed, such that redundant properties are encoded independently, then neural signals should not vary with complexity.
- Hypothesis 3: to decode such compressed representation, an active decompression operation should take place, with variable delays depending on the complexity and syntactic depth of the property that is being read-out.

Otherwise, if neural representations are factorized then accessing each property of the memorandum should take approximately the same time.

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

C. Results

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

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

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

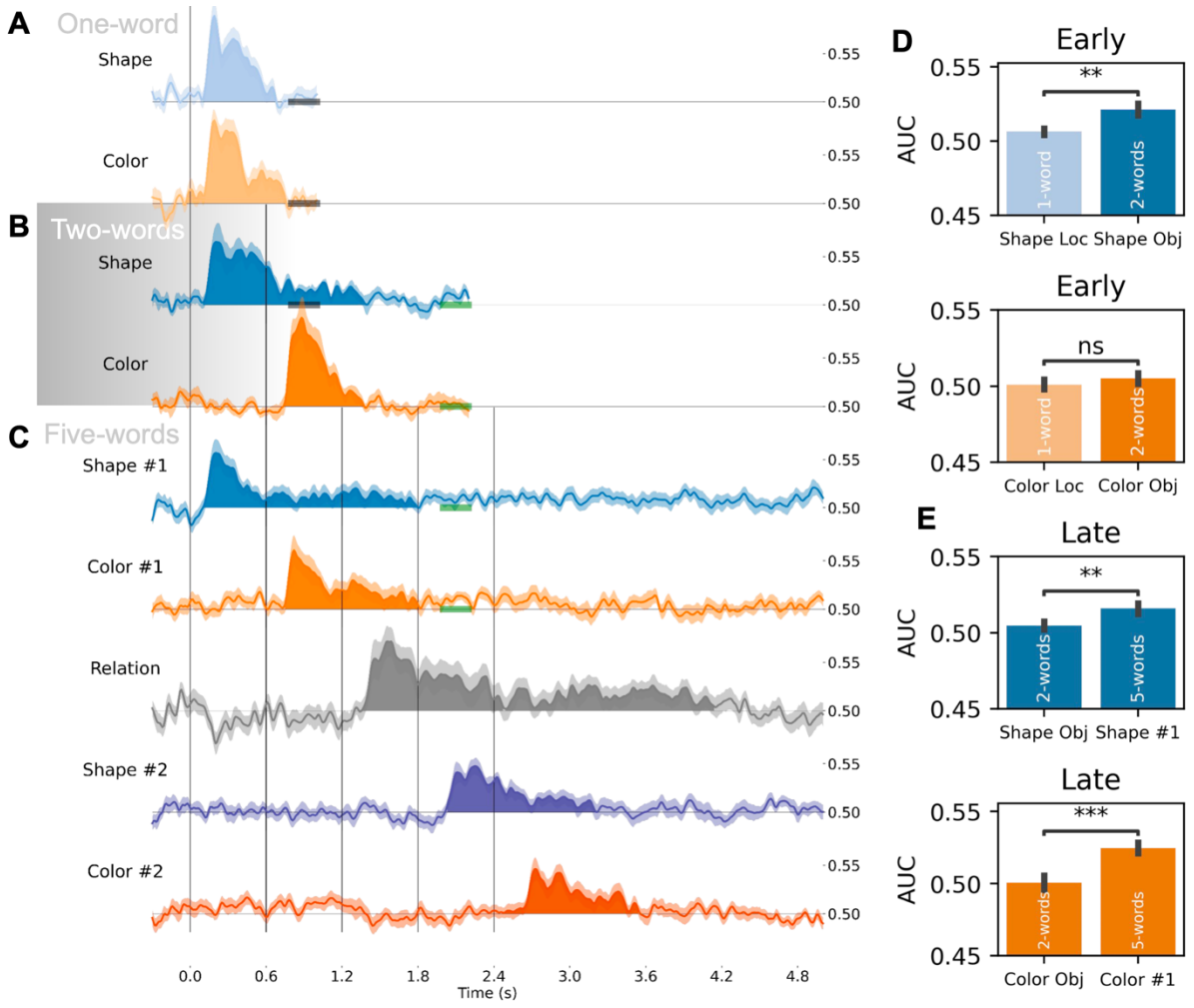


Figure 2-2: Single properties are actively represented until composition can occur

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

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

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

D, E: Evidence for sustained activity associated with phrasal composition. D, comparison of decoding performance between one-word blocks (left), where composition does not occur, and two-word blocks (right). All time points in the time window from 0.8 s to 1 s after word onset (thick black lines in panels A and B) were fed to the classifier. The decoding performance for shape in two-word blocks is higher than the one in one-word blocks, suggesting that the representation is kept active for composition. Using the same window for the second, color word (where in both cases no further composition should occur) does not yield any statistically significant difference. Statistics are FDR corrected.

E: Same for two-word blocks compared to five-word blocks, using a later time window (2 s to 2.2 s; thick green lines in panels B and C)), where composition should be completed in two-word blocks, but not in five-word blocks. For both shape (top, blue) and color (bottom, orange), the decoding performance is higher on five-words than on two-word blocks.

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

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

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

If the individual properties (shape, color, and relation) are stored in a way that is not detectable in MEG signals, what parameters impact the delay activity? We hypothesized that if the sentence representation is compressed, then the amount of information (Shannon, 1948) should be reflected in the ongoing neural activity. To test this, we devised a measure of complexity (C) that quantifies the information present in our sentences as the number of non-redundant objects' properties. Put simply, if both objects share color and shape, then C = 0, if they share either color or shape then C = 1, and if they share neither shape nor color then C = 2. We thus trained a linear regression to predict the complexity of

the composed representation of individual trials and evaluated the decoders with a Pearson correlation. Contrary to color and shape decoding where each class is predicted by a different classifier, this analysis learns a single mapping that predicts complexity values.

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

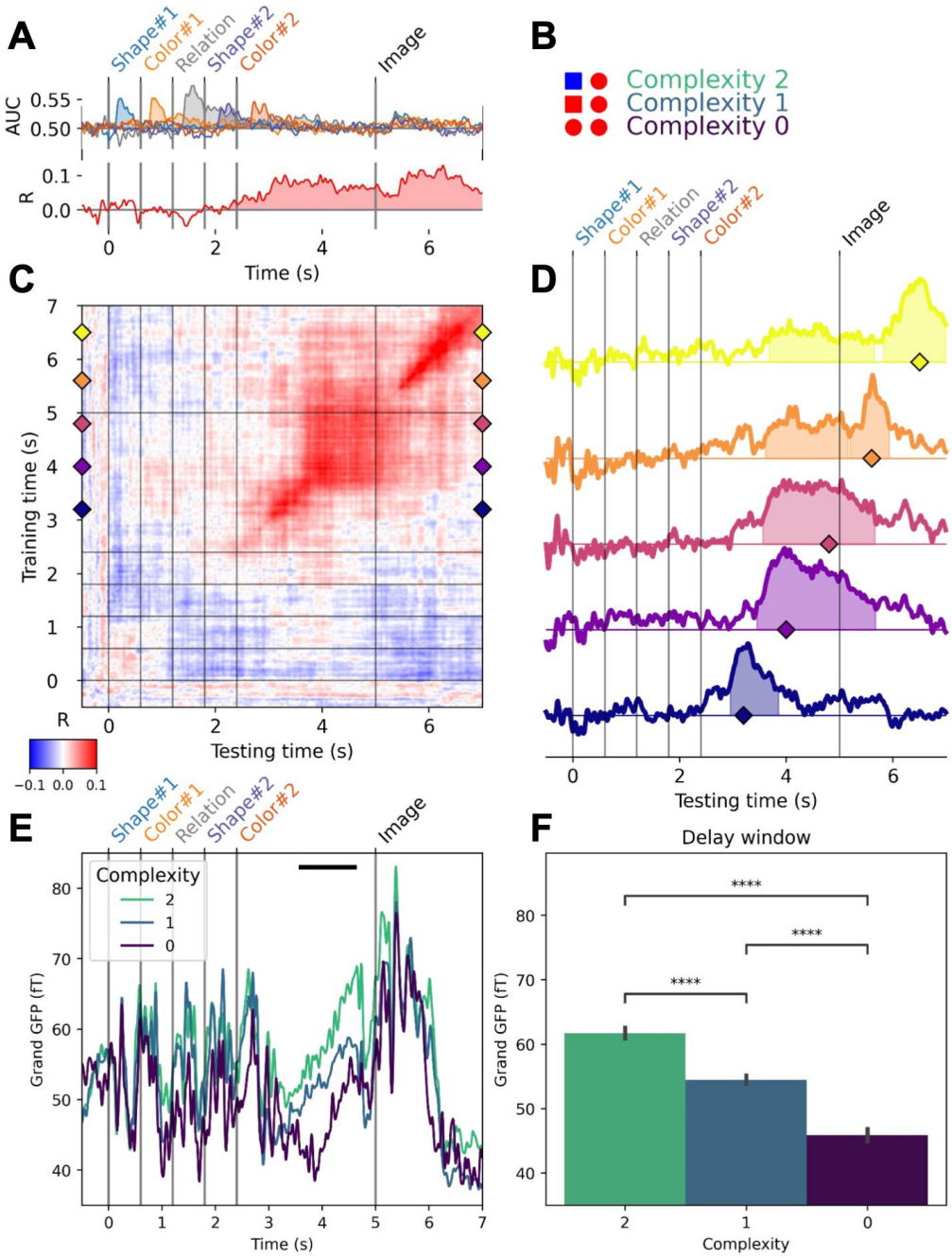


Figure 2-3: Impact of semantic complexity on neural activity in the delay period

A: Diagonal performance for decoders of individual word properties (top) and complexity (bottom). The feature decoders drop to chance level during the delay, precisely when complexity decoding becomes significant.

B: Example of trials for each complexity level, defined as the number of non-identical properties between the two objects.

C: Temporal generalization matrix for decoding of complexity

D: Horizontal slices through the temporal generalization matrix. Each graph shows the generalization performance of a fixed classifier, tested on MEG data from different time points. Early delay period

classifiers (e.g., 3 s, blue line) do not generalize well to later time points, whereas decoders trained later during the delay generalize up to and after the image probe (e.g., 4 s, purple line). Diamond markers represent the time each classifier was trained on.

E: Grand average of global field power (GFP) on magnetometers for each level of complexity.

F: Wilcoxon-Mann-Whitney test between each complexity level during the delay period (black bar in D). The data was averaged over all points in the time window and the test was performed over subjects.

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

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

3. Evidence for a decompression during read-out

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

To tackle this question, we focus on the neural activity between the presentation of the image probe and the subject's response. Firstly, we trained classifiers to differentiate trials where the probe corresponds to the preceding sentence (match) from trials where it does not (mismatch), for each complexity level. This gives us a window into the neural

processes that foreshadow the behavioral response. We expected that higher complexity trials should be associated with slower detection of mismatches. Indeed, we found that more complex trials are associated with later detection, both in neural signals (Figure 4A) and in reaction times (Figure 4B, right), but not in error rates (Figure 4B, left).

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

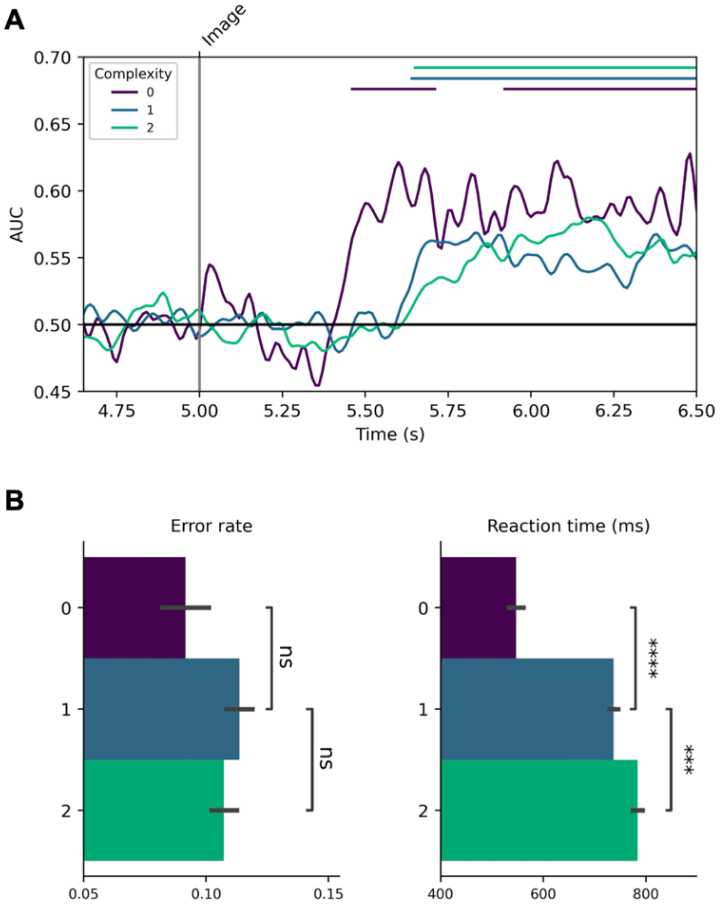


Figure 2-4: Effect of complexity at the time of image presentation

A: Decoding of match versus mismatch trials, split according to the three phrase complexity levels. Lower complexity is associated with an earlier separation of the brain signals for match and mismatch trials. The vertical bar marks image onset.

B: Average error rate (left) and response times (right) for each complexity level. Performance does not differ significantly with complexity, but response time is strongly affected.

Secondly, we take advantage of the hierarchical nature of our stimuli's syntactic trees, and the fact that in our experimental design, mismatch trials could be of three types. First, the *property* mismatches, where the shape or the color of one object was changed to a completely new one. Such a change in the surface properties of the syntactic tree is easily detected using a simple bag-of-words (Y. Zhang et al., 2010). Second, the *binding* mismatches, in which either the shape or the color of the two objects are swapped. To detect this mismatch, one needs to at least parse the noun phrases of the sentences, or in other words, correctly bind the objects' shape and color (Feldman, 2013). Third, the *relation* mismatches, where individual objects are preserved, but the spatial relation that links them is reversed (e.g., "X to the left of Y" as opposed to "X to the right of Y"). Thus, detecting this error requires reaching the uppermost branch of the syntactic tree and correctly assigning the objects' location (Figure 5 A).

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

Indeed, we found that the property mismatches were detected faster than binding and relation mismatches ($p<0.0001$ for both; Figure 5C right) and had a lower error rate compared to relation mismatches ($p<0.0001$; Figure 5C left). Comparing binding and relation mismatches, we find that the reaction times do not differ significantly ($p>0.05$; Figure 5C, right), but that the error rate is nearly twice greater for relation mismatches ($p<0.0001$, Figure 5C, left). This suggests a speed-accuracy trade-off (Reed, 1973) and corroborates the subjects 'verbal reports that strongly hinted that the hardest mismatches were the relation mismatches, and the easiest the property mismatches. Interestingly, many subjects reported that for relation mismatches in particular, they answered too soon and detected very soon after that they had made a mistake.

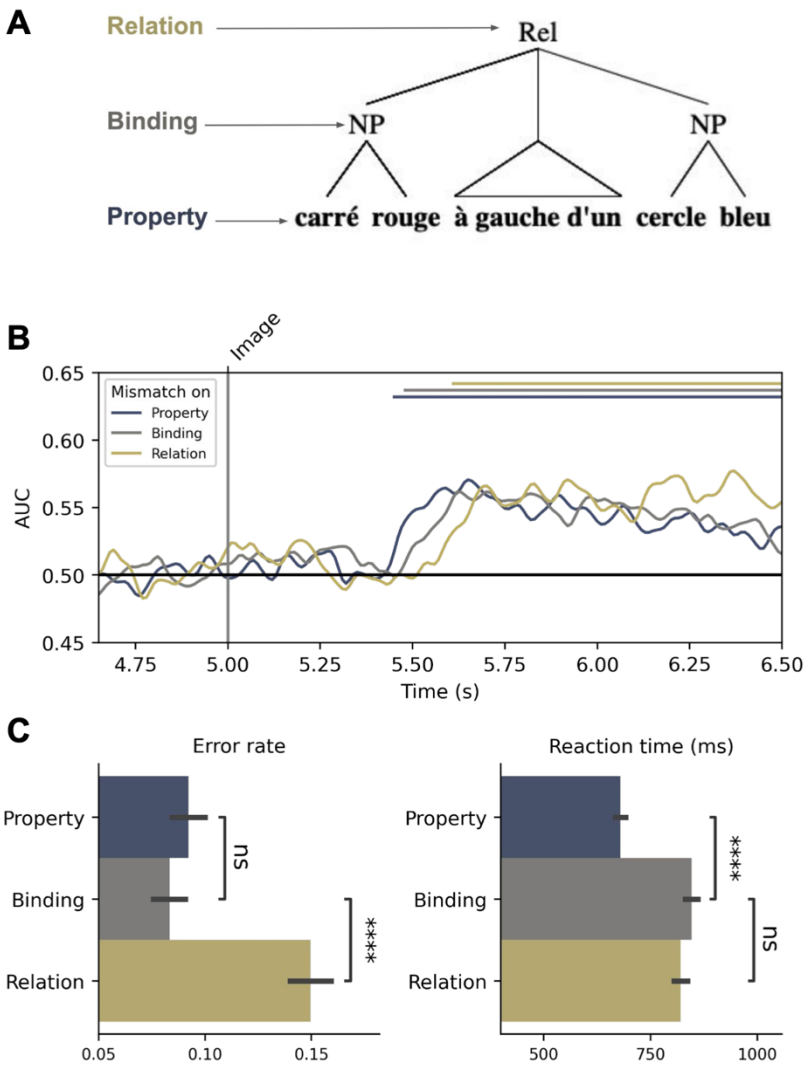


Figure 2-5: Effect of mismatch type at the time of image presentation

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

B: Decoding of match versus mismatch trials for each type of mismatch (Property, Binding, Relation). Increasingly delayed brain signals are observed for mismatches that concern deeper syntactic properties.

C: Average error rate (left) and response times (right) for each type of mismatch. Performance is lowest for Relation mismatches, while response time is highest for Binding and Relation mismatches.

We trained three separate sets of classifiers, one for each type of mismatch, and showed their respective performance on figure 5B. This decoding analysis confirms behavioral results, with the decoder trained on match versus property mismatch starting

first at 0.45 s after image onset, then the binding mismatch at 0.48 s, and the relation mismatch at 0.61 s.

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

D. Discussion

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

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

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

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

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

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

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

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

Note that there might be multiple alternative reasons why the properties decoding performance is at chance during the delay. First, the resolution allowed by MEG signals might not be sufficient if working memory is carried by sparse populations with sustained activity (Goldman-Rakic, 1995; Leung et al., 2002). Second, if working memory is

implemented in an activity-silent manner (Stokes, 2015), there might indeed be no way to decode its content without an external “probing” signal. Third, it might be that regular “replays” of the compressed word sequence are the basis of linguistic working memory, as has been shown for other sequences (Liu et al., 2019, 2021).

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

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

E. Methods

1. Experimental design

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

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

The experiment was split into 10 blocks, presented in a sandwich fashion: starting and ending with a one-word block, and alternating two and five-word blocks in between.

The 2 one-word blocks contained 480 trials each, totaling 960 trials. This means that for each of the 6 properties (3 shapes and 3 colors), we had 960 / 6 160 trials.

The 4 two-word blocks contained 135 trials each, totaling 540 trials. This means that for each of the object’ properties (shape, color), we had 540 / 3 = 180 trials. Thus, to train a classifier to differentiate trials where, e.g., the shape was 1) “square” versus 2) “circle or triangle”, we had 180 trials in the class 1 and 380 trials in class 2.

The 4 five-word blocks contained 81 trials each, totaling 324 trials. The number of unique sentences in our design is 3 first shape * 3 first_color * 2 relation * 3 second shape * 3 second_color = 162, thus we had 2 repetitions of each unique sentence. Furthermore, for each of the marginal objects’ properties (first shape, first color, second shape, second color), we had 324 / 3 = 108 trials. Thus, to train a classifier to differentiate trials where, e.g., the

first color was 1) “blue” versus 2) “red or green”, we had 108 trials in class 1 and 216 in class 2. For the spatial relationship property, we had 324 /2 = 162 trials in each category.

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

In two-word blocks, there was a single kind of mismatch: a new feature was selected at random to replace an existing one. E.g., “A blue square” becomes “A blue **circle***”.

Three kinds of mismatches were possible in mismatch trials (see also Supplementary Figure 1):

- In property mismatches, a new feature is selected to replace one, taken at random. This new feature could not already be present in the sentence. E.g.: “A blue circle to the left of a red square” becomes “A **green*** circle to the left of a red square”.
- In binding mismatches, two-words are swapped between the two objects. E.g.: “A blue circle to the left of a red square” becomes “A **red*** circle to the left of a **blue*** square”.
- In relation mismatches, the two objects are kept but the spatial relationship between the two is reversed. E.g.: “A blue circle to the left of a red square” becomes “A blue circle to the **right*** of a red square”.

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

2. Multivariate decoding

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

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

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

image were used, because a matching image would have confounded that memory trace with the incoming stimulus.

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

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

3. Global Field Power (GFP)

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

Contributions not included in the thesis

I was lucky to be part of scientific collaborations with many colleagues. Some of these resulted in published articles, enumerated in the "Publication of the author" section. Here, I would like to briefly summarize this work and my contributions to it.

First and foremost is the work on subject-verb agreement in artificial neural networks led by Yair Lakretz. In a series of papers, we assessed the performance of artificial neural language models on nested subject-verb agreement tasks as a window into long-distance dependencies. Basically, the task is to correctly conjugate the verb of a sentence, in conditions where it is separated from its subject by attractors of opposite number, e.g., "the **keys** to the cabinet **are** on the table" leads to more errors than "the **keys are** on the table".

We tested state-of-the-art recurrent neural networks (Lakretz et al., 2019; Lakretz, Desbordes, King, et al., 2021b), and transformers (Lakretz, Desbordes, Hupkes, et al., 2021; Lakretz et al., 2022) language models on this task and found that they systematically failed on some challenging constructions. Furthermore, we characterized a small network of specialized number and syntax units that seem to consistently emerge in recurrent neural networks trained on natural language (Lakretz et al., 2019).

In this work, my participation was mainly on the engineering side. Testing a large number of models on many tasks and including manipulations such as ablating each unit in a network one by one, is computationally challenging. I was tasked to generate scripts to run a large number of jobs on CPU and GPU cluster infrastructures, using the schedulers Slurm and Portable Batch System. The tools I used were mainly bash scripting, and the Python package submitit.

Second, the development of these tools led to the open-sourcing of our subject-verb agreement setup as part of the Beyond the Imitation Game benchmark (BIG-bench) (A. Srivastava et al., 2022). This set of tasks was made openly accessible, with the goal of assessing the abilities of current (and future) language models on difficult benchmarks.

There my contribution was minor, mainly adapting the existing code to fit the BIG-bench structure.

Third, along with fellow PhD students Christos-Nikolaos Zacharopoulos and Mathias Sablé-Meyer, we devised a new behavioral experiment based on subject-verb number agreement and tested it on human participants as well as state-of-the-art language models. Briefly, we sought to disentangle transition effects (i.e., surface-level statistics due to the close proximity of the attractor and target verb), and the structural effects (the mere presence of an attractor) that are classically confounded in such experiments. We find that both humans and models are sensitive to transition probabilities, but the latter more so. Importantly, we confirm that many previous studies confounded transition and structural effects.

Following an original idee from Christos, the three of us equally participated in the design and piloting. I was also responsible for neural network analyses, while Mathias took care of setting up the online experiment and most of the statistical tests. The writing of the paper was a joint enterprise, led by Christos. I will remember this collaboration as one of the most stimulating parts of the PhD. Our motivation, taking its roots in raw curiosity, the freedom that we were given, and the challenges that we had to face with (voluntary) limited interaction with our supervisors, jointly brought about an elegant study that we all enjoyed working on.

These three sets of studies with a focus on the subject-verb agreement task, do not directly tackle the question of compositionality. Rather, they present a more specific analysis of the mechanisms used to carry grammatical features during online sentence processing. As such, they provide a complementary view of the computations underlying language processing in brains and machines.

General discussion

A. Summary of the main results

Compositionality long seemed to be very difficult to study experimentally, but today it seems that this "holy grail" of cognitive science (R. Jackendoff, 2002) is no longer out of reach. Researchers from many different fields are coming together to try to unravel the mysteries of an ability that makes humans so special. In this thesis, I sought to identify neural correlates of meaning composition, borrowing tools from neuroscience, linguistics, and artificial intelligence. I presented two studies that provide complementary views on the subject.

Let's start this discussion by bringing together the main results of the two studies and organize them along two central pivots: temporal dynamics and format of representations.

The first goal of our work was to characterize the temporal dynamics of semantic composition. This was the main focus of the first study, in which we presented an overview of the three dynamical processes happening when multiple words have to be integrated with one another, namely lexical access, multi-word integration, and final wrap-up. We identified varying degrees of participation in each process for different brain regions, with a leading role of the frontal cortex in compositional operations. The second study also provides a valuable insight into the temporal dynamics of composition. There, we managed to decode each word in phrases of varying length and discovered that words are maintained longer in neural activity when they are part of longer phrases. Taken together, these temporal dynamics suggest a delayed composition model, wherein each word is maintained until it can be meaningfully integrated with its larger context.

The second goal was to analyze the format of compositional representations. In the later study, we found that compositional representations in working memory are compressed. Additionally, read-out from working memory was associated with task-specific computations, suggesting that transformations of the neural representations were necessary to access the full syntactic tree of the sentence. In the first study, we also

introduced a measure of intrinsic dimensionality to the field of neurolinguistics and showed that meaningful composition was associated with higher dimensionality. These results suggest that the intrinsic dimensionality of a neural representation is a good index of its semantic content, possibly portraying the amount of incompressible information.

Overall, this thesis provides new insights into the dynamics and structure of linguistic composition. Let's now discuss the limits of this work and the prospects for future work.

B. Limitations and future work

The present effort could (and hopefully will) be complemented in many ways. The temporal dynamics results complement a growing body of literature that sought to characterize the transitory neural activity happening during composition (Hagoort, 2019; Pylkkänen, 2019). Comparatively, there is little material on the format of compositional representations. This section discusses the limitations of the present studies and proposes directions for future work with a focus on this last target.

First regarding work in the making, the paradigm of the second study will be adapted to functional magnetic resonance imaging (fMRI). This modality is especially well suited to localize the effects we observe, thanks to its excellent spatial resolution. Indeed, though MEG provides millisecond-level temporal resolution and is, therefore, ideal for measuring rapidly unfolding language processing, its spatial resolution is orders of magnitude coarser than fMRI's. This has important consequences regarding what kind of information can be identified on these signals. A recent study has found compositional representations in fMRI, but not MEG (Toneva et al., 2022). One outstanding analysis that we failed to set up in the second study, would be to successfully decode mirror-image sentences ("X to the left of Y" and "Y to the right of X"), that carry the same meaning through different word sequences. We could thus identify regions that encode compositional meaning irrespective of word order.

Second, using the data acquired during this thesis or an fMRI follow-up, training encoding models could be a great way to further characterize the neural code of

compositional representations. Encoding models predict brain activations from experimental variables, thus allowing to identify the main parameters that influence neural activity in each voxel or sensor. This could help answering questions such as: how are each word's features combined when composition occurs, i.e., are the word representations summed, multiplied, or some other operation? Are nouns, verbs, and other parts-of-speech, represented in spatially distinct subregions?

Third, our results will need to be reproduced in more naturalistic setups (see below the section on the debate between controlled versus naturalistic experimental setups). This should prove especially difficult, because the methods we used necessitates many repetitions of each word, both in isolation and in increasingly longer phrases. Leveraging the big data made available by large open-source datasets (Nastase et al., 2020; Allen et al., 2022; Armeni et al., 2022) could be a promising though challenging avenue for future work. Of special interest, it would allow to test if the representation of a sentence is reactivated when subsequent sentences refer to it.

Fourth, I speculate that the intrinsic dimensionality analysis has a lot of potential to unravel compositional representations both in the brain and in artificial neural models. Indeed, this method is very general, and well suited to characterize how the neural space is populated. In addition, it does not presuppose a particular theory of sentence composition. Therefore, applying this method to new paradigms and new datasets could prove a fruitful endeavor.

Fifth, another program of research could focus on language production. There, the processes are reversed compared to language comprehension: going from a (compositional) concept, one has to generate the appropriate sequence of words to convey its meaning. It is possible that we would see similar processes to the ones we saw when extracting a representation from working memory, i.e., a kind of "decompression" when reading semantic content from long-term memory.

To sum-up this section, the limits specific to the present studies can be solved in future experiments. Several such extensions are discussed. Let us now discuss more general limitations of our work.

C. General limitations

Several limitations peculiar to each study are described above, as well as in their respective chapters. Here, I'd like to discuss what is, in my eyes, the main drawback of this thesis, and indeed much of the neuroimaging literature. That is, that our results are correlational in nature, and as such, they only inform us about the properties of the system under study, but not so much about the precise calculations that are done. In other words, some of the neural signatures that are identified (in both space and time) could be the byproduct of processes of a different nature than the one we hypothesize.

Here, we haven't yet causally tested hypotheses regarding the actual computations that are done to combine words. Our results provide a temporal and spatial overview of the processes happening during composition, as well as some attributes of compositional representations, but we haven't yet identified the exact operations that take place.

One way out of this issue is through advances in neuroimaging techniques, a thriving area (Seo et al., 2016; Bihan & Schild, 2017; Musk & others, 2019; Steinmetz et al., 2021). The ability to precisely record larger neural populations would provide us with opportunities to test more precise hypotheses about information encoding and processing in the brain. Crucially, the ability to both record and stimulate would allow to make causal experiments and thus go beyond mere correlations.

Another way out is through theoretical work. There are many theories of language composition, coming from linguistics and artificial intelligence, and they can be used to derive testable predictions. In a section below, I discuss some of these theories, and how our results relate to them.

D. Implications for theories of sentence composition

Here I would like to describe some of the main theories of linguistic composition in the brain, their tenants, and the implications our results have in their regard.

First, vector-symbolic architectures, such as the tensor product representation (Smolensky, 1990) and holographic reduced representation (Plate, 1995) have already been introduced in preceding chapters, but I would like to take the time to discuss them in a more detailed fashion. Put simply, they propose that the encoding of meaning is distributed in very large populations of neurons. This view is opposed to more classical "localist" views, where meaning is represented, in the extreme case, in a single neuron such as a grandmother cell (Quiroga et al., 2005).

These proposals take advantage of the computational power of vector arithmetic operations and the properties of high-dimensional spaces, two postulates that reasonably apply to the brain. An example of such properties is that, in high-dimensional spaces, random vectors are guaranteed to be quasi-orthogonal, making straightforward the storage of new elements independently of each other. By way of explanation, a vector space has a number of strictly orthogonal directions equal to its dimensionality, but the number of quasi-orthogonal directions increases exponentially with the dimensionality (this is sometimes called the blessing of dimensionality (Gorban et al., 2020), in opposition to the curse of dimensionality in machine learning). Additionally, the similarity between two representations is easily computed in such vector spaces, e.g., by means of dot-product. Without going further in the details, vector-symbolic architecture solve the issues of early connectionist systems raised by Fodor and Pylyshyn, namely productivity, systematicity, compositionality, and inferential coherence (Fodor & Pylyshyn, 1988; Kleyko et al., 2022)

In tensor-product representation (Smolensky, 1990), two kinds of vectors are combined to represent the meaning of a word in a particular context: the "filler" vector contains the meaning of the word (its lexical semantics), while the "role" vector encodes its syntactic role in the sentence. These vectors are randomly sampled across the vector space to construct the "dictionary". These atomic vectors are combined by means of an outer-product (also called tensor-product). To encode the meaning of a sentence, this operation is applied to each word and the resulting tensors are summed. This final tensor contains the compositional meaning of the sentence. Individual elements (role or filler atomic vectors) can be recovered by means of tensor-vector inner product, where the query vector is computed from the "dictionary" matrix storing each atomic vector.

Holographic reduced representation (Plate, 1995), sometimes also called Semantic pointer architecture (Eliasmith et al., 2012), is a variant that solves the exponential growth of dimensionality of tensor-product representation. Indeed, the recursive application of the outer-product yields tensors of ever-increasing dimensionality, weakening their biological plausibility and practical use. Holographic reduced representations, however, compress the compositional tensors back into the dimensionality of the original vectors by means of a circular convolution. This operation approximately perpetuates the norms of the input vectors, thus preserving most of their information, despite relying on a fixed-length vector to store compositional meaning. Atomic vectors can be recovered using circular correlation with the query vector, followed by a nearest neighbor search in the dictionary.

Although our results do not directly confirm or invalidate these theories, interesting parallels can be made especially with the former. Most strikingly, the compression of compositional representations in working memory could correspond to the "reduction" operation in holographic reduced representations. Taking this view, the earlier phase of increased complexity decoding could correspond to the outer-product between the role and filler vector of each word, or to the summation of each role-filler tensors, just before compression. Indeed, in this framework, a natural prediction is that before summation and compression, the role-filler tensors associated with each word would still linearly represent the word's characteristics, but after summation and compression, this information would not be linearly accessible anymore.

Parallels can also be made with theories of the composition of morphemes into words. A recent model (Gwilliams, 2020) proposes, that during language comprehension, four stages happen to identify the meaning of a word: i) segmentation, i.e. the identification of the morphemes that are present, ii) look-up, where the morphemes are linked to their respective semantic or syntactic features, iii) composition, the combination of features into a complex representation, and iv) update, an operation changing the rules and features, if necessary.

Interesting links can be made with multi-word compositions. One is led to ask if and how these stages could map to the three stages we described in the first study: lexical access, multi-word integration, final wrap-up. It could be that a first stage is missing and would actually correspond to the stage of morphemic composition: identifying the word, accessing its meaning, composing it with other words, double-checking or updating the rules of composition. Crucially, in Gwilliams' proposal, the stages are not strictly sequential, but are allowed to overlap in time. Indeed, in our case, it is likely that an incoming word is undergoing lexical access, while the preceding ones are still being integrated with one another.

Last, the recent success of deep learning for natural language processing provides us with valuable tools to study compositional representations in the brain. Taken to the extreme, one could view contextual word embeddings as the instantiation of a particular theory of word composition, although one that cannot be articulated with words. The main components of these theories would be the architecture of the model, its objective and cost functions as well as the dataset it was trained on. Such computational theories can be compared by way of linear encoding models to brain activations (Hale et al., 2022), and the most important components identified in a systematic way (Schrimpf, Kubilius, Lee, et al., 2020; Caucheteux et al., 2021b; Caucheteux & King, 2022; Pasquiou et al., 2022). This use of neural language models as models of the brain, in opposition to more classical linguistic theories, is discussed in the section below, as well as other controversies present in the field.

E. Debates in the field

The field of neurolinguistics, and neuroscience more generally, is a lively place regularly shaken by stimulating debates. Here I describe three of them: i) controlled versus naturalistic designs, ii) the use of neural language models as models of language in the brain, and iii) should artificial intelligence take inspiration from neuroscience, and discuss this thesis' stance in their regard.

Classically, all experimental sciences have worked with tightly controlled experimental conditions (Fisher, 1936), with the hope that conclusions would generalize to broader contexts. And in many cases, they did, as exemplified in the famous quote by molecular biologist Jacques Monod “What is true for the bacterium must be true for the elephant”. But does it hold for neurolinguistics? Is it accurate that what is true in sentences versus word lists must be true in a real-life conversation? In other words, can the findings from highly controlled stimuli in a lab environment generalize to natural, authentic behavior?

This reductionist approach (in the sense that it decomposes a difficult puzzle in a series of simpler, fundamental problems) relies on the postulate that the cognitive functions under study can be decomposed and put back together and still yield a satisfying description of neural dynamics and behavior. It has been recently argued that tightly controlled linguistic stimuli could yield conclusions that are specific to the task at hand, and would not generalize to other linguistic tasks (Varoquaux & Poldrack, 2019). In other words, the risk is, as Brunswik warned more than half a century ago, to confine ourselves in “a self-created ivory-tower ecology” (Brunswik, 1956, p. 110). In response, some researchers proposed that we should focus on naturalistic paradigms, such as audiobooks or movies (Hamilton & Huth, 2020; Willems et al., 2020).

In vision neuroscience, this question arose some 20 years ago. For example, it was found that the receptive fields of neurons in primary visual areas are altered when viewing complex stimuli (David et al., 2004), compared to the canonical receptive fields (Hubel & Wiesel, 1962). More recently, the existence of face selective regions in humans (Kanwisher et al., 1997) and macaques (Tsao et al., 2006) have been put into question. Indeed, when presented with natural stimuli, the regions showed more varied behavior, with a dependence on agentic action in humans (Haxby et al., 2020), and spatial scale and social scene in macaques (McMahon et al., 2015). Certainly, controlled experimental setups were required to detect these regions in the first place, but they failed to provide a complete picture of the role of these regions.

It is undoubtable that the foundation of our field of study is built on such controlled experiments where few conditions are isolated and contrasted. It is easy to understand why:

the immense complexity of the brain and high variability in neural recordings and in behavior prompted researchers to make simple, tightly controlled experiments to get reproducible findings. But ultimately, understanding the brain will require putting back together pieces of knowledge from many different experiments, and naturalistic paradigms can assist us in this goal. It is still too early to tell which findings will pass the test of time. Most likely, the main findings should hold, but provide an incomplete picture of language processing in the brain. For example, the semantic networks were found to be much larger than previously thought when naturalistic stimuli were used (Huth, Lee, et al., 2016).

In this thesis, we take a stance somewhat in between the two extremes of complete control or naturalism. In the first study, the sentences used are based on a large vocabulary, and thus relatively natural. The use of neural language models, trained on natural text, to validate our hypotheses, also supports the generality of our findings. In the second study the stimuli used were comparatively more controlled, using a smaller vocabulary, which allowed us to test more precise hypotheses about the storage of compositional representations.

Going forward, what is the best course of action for a junior neuroscientist? First, it is always good to keep in mind the context to which our hypotheses apply (Holleman et al., 2020). More general findings should, in fine, be validated in more naturalistic paradigms. Second, naturalness and well controlled parametric design are not necessarily incompatible and should be combined whenever possible. Some recent studies have managed to get the best of both worlds. For example, Macdonald and Tatler varied the amount of gaze cueing in the oral instructions for an interlocking toy bricks task, finding that gaze cues led to more accurate performance (Macdonald & Tatler, 2013). Receiving instruction to build a structure is certainly natural, surely most of us (tried to) build some unassembled furniture. In another outstanding example, Goldstein and colleagues (Goldstein, Zada, et al., 2022) showed that humans naturally engage in next-word prediction during natural story listening. This provides a behavioral explanation for the previously unexplained success of neural language models in predicting brain activity during language comprehension. A related approach is to take natural stimuli and manipulate them to test specific hypotheses. For example, taking natural stimuli and degrading them by adding noise to study how phonemic representations are altered (Di Liberto et al., 2018); or removing harmonics in the natural

speech signals impairs intelligibility and followability (Popham et al., 2018). These should serve as examples that careful manipulations can be done in naturalistic settings.

To conclude on this, I surmise that we should trust in the incremental principles at the heart of science: both controlled and natural setup can yield compelling, complementary insights.

Let's turn to another heated debate: do current artificial intelligence algorithms provide good models of the brain? Historically, cognitive science and artificial intelligence have been developed in tandem, with many pioneering work that can't really be attributed to one field or the other, but rightfully belongs to both (McCulloch & Pitts, 1943; Hebb, 1949; Turing, 1950; Marr & Poggio, 1976; Sutton & Barto, 1981; Hopfield, 1982; Hinton, 1984; Rumelhart et al., 1985; P. S. Churchland & Sejnowski, 1988). The link was weakened in following years, but it might be coming back following the recent successes of deep learning models, both in traditional natural language processing (NLP) benchmarks, and as predictive models of brain activity (Yamins & DiCarlo, 2016; Hale et al., 2022).

State-of-the-art neural language models are known to be (at least somewhat) biologically implausible: they are trained with backpropagation (Rumelhart et al., 1985), have some architecture that do not match to low-level neural processes and brain-level organization principles, such as recurrence (Gwilliams & King, 2020). However, their representations are currently the best predictors of brain activations in a wide range of tasks, as described in the introduction of chapter 1. Curiously, even untrained language models yielded above-chance predictivity of brain activations, with recurrent neural networks performing better than transformers, a trend that reverses during learning (Pasquiou et al., 2022). Is there something special about these architectures that makes them brain-like? Strikingly, it was recently found that, under biologically-plausible assumptions, the transformer architecture is equivalent to models of place and grid cells in the hippocampal formation (Whittington et al., 2022). This suggests that at the computational level, such artificial algorithms can actually be good models of cerebral mechanisms. Additionally, in a recent study researchers considered intracranial recordings

in the inferior frontal gyrus as brain-embeddings, and compared them to contextual word embeddings from neural language models (Goldstein, Dabush, et al., 2022b). They found that they have a similar geometry, suggesting that the brain uses its neurons somewhat like a contextual word embedding. These surprising convergences provide a new explanation for the ability of such language models to predict brain activations, as well as a fresh view on the format of biological linguistic representations.

Let's now consider neural networks not just as predictive models of brain activations, but as actual models of the brain, as was suggested by some researchers (Richards et al., 2019; Hasson et al., 2020). Examples of successes in this regard include networks trained on natural tasks that exhibit activation patterns such as grid cells (Banino et al., 2018), temporal receptive field (Singer et al., 2018), and behavior such as model-based reasoning (Wang et al., 2018; Wang, 2021).

Moreover, the implausibility of error backpropagation, one of the main argument against the use of artificial neural network as models of the brain, has been put into question recently, as biologically plausible alternatives have been discovered (Lillicrap et al., 2016; Guerguiev et al., 2017; Whittington & Bogacz, 2017; Bellec et al., 2020; Illing et al., 2021). This should alleviate some of the doubts that neuroscientists could have regarding the plausibility of these models, stemming from the huge implementational differences between brains and such models.

In addition, arguments against the over-parameterization of such networks have arisen, but these are again put into question, as over-parameterization is seen by some as a necessary condition for generalization behaviors of ANNs (Neyshabur et al., 2018; Hasson et al., 2020; Advani et al., 2020). Furthermore, the brain seems vastly over-parametrized as well (or at least, vastly more complex than current deep learning models). It is likely that to fully explain natural neural responses, one would need a compact framework that allows for fitting billions of parameters to match the complexity of the brain (Richards et al., 2019). I speculate that the best description of the brain, if such is possible, would not be in the form of words, but rather in the form of a small set of mathematical principles, somewhat like the ones used in neural networks: architecture, objective function and learning rule.

On a side note, irrespective of the debate of whether deep learning can provide good brain models or not, it is undeniable that it brings new valuable tools to neural data analysis and novel experimental designs. For example, artificial neural networks were used to generate images that maximally activates spiking activity of single neurons (Bashivan et al., 2019), providing new insights into biological neuron sensitivity.

To sum up, as time goes by, artificial neural networks are getting more readily accepted as brain models, and their use as tools for neural data analysis is increasing. This makes them a crucial tool for the future of neurolinguistics.

Let's now consider the converse question: should artificial intelligence (AI) take inspiration from neuroscience to develop new algorithms?

Many researchers have argued that taking inspiration from neuroscience and cognitive science is a promising avenue for AI (Hassabis et al., 2017; Lake et al., 2017; Marcus, 2020; Bengio et al., 2021). In this regards, success stories abounds: dropout mimicking the unreliability of neuronal discharge (Hinton et al., 2012; N. Srivastava et al., 2014; Fan et al., 2019), divisive normalization and maximum-based pooling of inputs (Carandini & Heeger, 2012; Yamins & DiCarlo, 2016; Sanchez-Giraldo et al., 2019), and attentional mechanisms (Bahdanau et al., 2014; A. Graves et al., 2016; Vaswani et al., 2017), just to name a few.

Some insist that, despite their good performance at their artificial tasks, artificial neural networks are still far from human-, or even animal-level intelligence (Zador, 2019), and could learn from them. One argument is that a lot of animal behavior is innate, for example, spiders are ready to hunt as soon as they are born. Similarly, in the domain of language, it has been argued that infants are exposed to too little data to learn language from scratch (Chomsky, 1980). This so-called poverty of stimulus argument, described in more detail in the introduction, suggests that there is some innate machinery that guides language learning in children. This brings us back to the classic nature versus nurture debate.

By contrast, current artificial neural networks are sometimes argued to be a blank slate, or *tabula rasa* before training (Silver et al., 2017; Hahn & Baroni, 2019; Jaderberg et al., 2019). However, some researchers view the choice of i) architecture, ii) objective function, and iii) learning rule as major prior knowledge put into the system (Richards et al., 2019). One could argue that the choice of training dataset is also part of the prior knowledge, as it has been shown to influence the representation of trained networks (Pasquiou et al., 2022). These four components, crucial to the final behavior of the network, are left to the experimenter to choose. Curiously, previous symbolic-based AI, where many decisions were left to the researcher (Newell & Simon, 1961), never reached a level of performance similar to current deep learning based methods.

One view on this suggests that gradient-based training (both supervised and unsupervised) could be seen as an analog to biological evolution rather than individual learning (Zador, 2019). In this sense, the enormous amount of data needed to feed the model would be matched by the gargantuan number of individuals, and all their experiences, in the history of a species. Thus, the relatively small number of genes (many orders of magnitude fewer than the number of synapses in the brain!) in the genome could be seen as a "regularizer" (Poggio et al., 1987; Bickel et al., 2006), or an information bottleneck (Tishby et al., 2000; Tishby & Zaslavsky, 2015; Saxe et al., 2019), shifting the balance from variance to bias (Geman et al., 1992) and avoiding overfitting. Relatedly, it is worth mentioning that only general structural and wiring rules are selected by evolution, as anything more precise would require more storage than what is available in the genome. From these wiring rules, development and learning must actualize appropriate behavior. This suggests that the secret recipe to human intelligence should be found in network wiring and topologies. Indeed, game-changing papers in machine learning often involve a new architecture (LeCun et al., 1989; Goodfellow et al., 2014; Kingma & Welling, 2014; Devlin et al., 2019; Vaswani et al., 2017). One could find this surprising, given that artificial neural networks are universal function approximators (Cybenko, 1989; Hornik, 1991), even in the "vanilla", fully-connected flavor. What to take from this is that these architectures have inductive biases (T. M. Mitchell, 1980; Richards et al., 2019) that guide learning and allow better generalization. This discussion is further muddled by a difference in appellations: linguists and psychologists speak of "constraints", and machine learning scientists of

“inductive biases”, while statisticians and bayesians from various fields keep using the term “priors”. Hopefully, they all allude to similar meanings.

So where does this thesis stand? Obviously, there is no absolute right or wrong here. For intelligence to thrive, there should be both nature and nurture. Actually, most interesting behaviors emerge from the interaction of the two. For example, so-called “place” and “grid” cells found in the hippocampus and entorhinal cortex of most mammals (reviewed in (Moser et al., 2008)). They were first thought to represent a “simple” tessellation of space that allows to compute the current location of the animal. They have since been shown to be attuned to cognitive factors such as rewards and goals (Boccara et al., 2019), such that they enact a cognitive map (Behrens et al., 2018) of the immediate spatial environment. This faculty is innate: when a rat pup explores an open environment for the very first time, such a map of space quickly emerges (Langston et al., 2010). But the content of the map is learned and highly dynamic: if the known environment of an adult rat changes a bit, the places cells will undergo a “remapping” to reflect this change (Cressant et al., 2002; Fyhn et al., 2007). Similarly for language, there might be some specialized innate structure, but undeniably it is interacting with the environment that allows a full development of linguistic ability. The scaffolding may be innate, but the content built on this scaffolding is learned.

How to formally characterize this interplay of evolution and learning in an individual? An elegant framework that could explain this is meta-learning (Andrychowicz et al., 2016; Bellec et al., 2018; Wang et al., 2018; Wang, 2021; Hospedales et al., 2022). In meta learning, an outer loop selects the best hyper-parameters for a family of learning algorithms over the course of many “episodes”, or inner loops, during which limited learning occurs. Basically, the outer loop optimizes for general properties of the task (or tasks), while the inner loop adapts to the peculiarities of the particular episode. Thus, one could see evolution as an outer loop that optimizes the learning algorithms and inductive biases of the individuals, who would then be well equipped to appropriately learn during their lifetime.

Back to machine learning, in language modeling, the presentation of many sequences, with no gradient update inside a sequence (inner loop) but only between sequences (outer-loop), can also be seen as meta-learning. Some researchers think this

explains the impressive generalization skills of modern neural language models (Radford et al., 2019; Brown et al., 2020b).

In this section I have explored debates regarding naturalistic versus controlled designs, as well the interplay between neuroscience and AI. It is time to conclude.

F. The end goal of neurolinguistics?

If the end goal of neurolinguistics is to achieve a thorough understanding of language processing in the brain, then under what condition can this goal be fulfilled? When can we say that we have reached a good-enough understanding and be satisfied? I contend that a reasonable, though challenging goal would be to truly bridge the fields of linguistics and neuroscience. Specifically, to describe direct correspondences between the jargon used by scientists of both fields, the so-called Mapping Problem between linguistics and neurobiology (Embick & Poeppel, 2015; Hale et al., 2022). For now, it is clear that a systematic correspondence is missing: there is no linguistic concepts that directly maps to neuroscientific notions such as “spike train”, “oscillation”, and “population vector”; and conversely there is yet no neuroscientific counterpart to linguistic notions like “noun phrase”, “movement”, or “merge”. A potential way towards that goal is to embrace natural language processing tools. They have recently been put to use by neuroscientists eager to find the models that best predict brain activity but could be used more generally as a middle ground between the two fields. It may be easier to map characteristics of such behaving models to notions in linguistics and neuroscience than it would be with a direct mapping. For example, the classical word embeddings, such as word2vec, straightforwardly map to the linguistic notion of lexical semantics, and can be used as features for brain encoding models, thus providing a long sought-for bridge. I anticipate that such links will flourish in years to come.

Bibliography

Advani, M. S., Saxe, A. M., & Sompolinsky, H. (2020). High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, *132*, 428–446. https://doi.org/10.1016/j.neunet.2020.08.022

Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2022). Negative sentences exhibit a sustained effect in delayed verification tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(1), 122–141. https://doi.org/10.1037/xlm0001059

Agrawal, A., Hari, K., & Arun, S. P. (2020). A compositional neural code in high-level visual cortex can explain jumbled word reading. *ELife*, *9*, e54846. https://doi.org/10.7554/eLife.54846

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

Alderson-Day, B., & Fernyhough, C. (2015). Inner Speech: Development, Cognitive Functions, Phenomenology, and Neurobiology. *Psychological Bulletin*, *141*(5), 931–965. https://doi.org/10.1037/bul0000021

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial

intelligence. *Nature Neuroscience*, *25*(1), 116–126. https://doi.org/10.1038/s41593-021-00962-x

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B., & De Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. *Advances in Neural Information Processing Systems*, *29*.

Armeni, K., Güçlü, U., van Gerven, M., & Schoffelen, J.-M. (2022). A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. *Scientific Data*, *9*(1), Article 1. https://doi.org/10.1038/s41597-022-01382-7

Arnauld, A., & Lancelot, Claude. (1660). *Grammaire générale et raisonnée...* Paris Durand.

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

Badier, J. M., Dubarry, A. S., Gavaret, M., Chen, S., Trébuchon, A. S., Marquis, P., Régis, J., Bartolomei, F., Bénar, C. G., & Carron, R. (2017). *Technical solutions for simultaneous MEG and SEEG recordings: Towards routine clinical use*. *38*(10), N118–N127. https://doi.org/10.1088/1361-6579/aa7655

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *ArXiv Preprint ArXiv:1409.0473*.

Bancaud, J., Angelergues, R., Bernouilli, C., Bonis, A., Bordas-Ferrer, M., Bresson, M., Buser, P., Covello, L., Morel, P., Szikla, G., Takeda, A., & Talairach, J. (1970). Functional stereotaxic exploration (SEEG) of epilepsy. *Electroencephalography and Clinical Neurophysiology*, *28*(1), 85–86.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., Modayil, J., Wayne, G., Soyer, H., Viola, F., Zhang, B., Goroshin, R., Rabinowitz, N., Pascanu, R., Beattie, C., Petersen, S., … Kumaran, D. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429–433. https://doi.org/10.1038/s41586-018-0102-6

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

Baron, S. G., & Osherson, D. (2011). Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage*, *55*(4), 1847–1852. https://doi.org/10.1016/j.neuroimage.2011.01.066

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

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science (New York, N.Y.)*, *364*(6439), eaav9436. https://doi.org/10.1126/science.aav9436

Bastiaansen, M., & Hagoort, P. (2006). Oscillatory neuronal dynamics during language comprehension. In C. Neuper & W. Klimesch (Eds.), *Progress in Brain Research* (Vol. 159, pp. 179–196). Elsevier. https://doi.org/10.1016/S0079-6123(06)59012-0

Bastiaansen, M., Magyari, L., & Hagoort, P. (2009). Syntactic Unification Operations Are Reflected in Oscillatory Dynamics during On-line Sentence Comprehension. *Journal of Cognitive Neuroscience*, *22*(7), 1333–1347. https://doi.org/10.1162/jocn.2009.21283

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

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

Bellec, G., Salaj, D., Subramoney, A., Legenstein, R., & Maass, W. (2018). Long short-term memory and learning-to-learn in networks of spiking neurons. *Advances in Neural Information Processing Systems*, *31*.

Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., & Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-17236-y

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

Bemis, D. K., & Pylkkänen, L. (2013a). Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral Cortex (New York, N.Y.: 1991)*, *23*(8), 1859–1873. https://doi.org/10.1093/cercor/bhs170

Bemis, D. K., & Pylkkänen, L. (2013b). Basic Linguistic Composition Recruits the Left Anterior Temporal Lobe and Left Angular Gyrus During Both Listening and Reading. *Cerebral Cortex*, *23*(8), 1859–1873. https://doi.org/10.1093/cercor/bhs170

Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep learning for AI. *Communications of the ACM*, *64*(7), 58–65.

Berger, H. (1929). Über das elektroenkephalogramm des menschen. *Archiv Für Psychiatrie Und Nervenkrankheiten*, *87*(1), 527–570.

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

Berwick, R. C., Chomsky, N., & Piattelli-Palmarini, M. (2013). Poverty of the stimulus stands: Why recent challenges fail. *Rich Languages from Poor Inputs*, 19–42.

Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the Stimulus Revisited. *Cognitive Science*, *35*(7), 1207–1242. https://doi.org/10.1111/j.1551-6709.2011.01189.x

Bickel, P. J., Li, B., Tsybakov, A. B., van de Geer, S. A., Yu, B., Valdés, T., Rivero, C., Fan, J., & van der Vaart, A. (2006). Regularization in statistics. *Test*, *15*(2), 271–344. https://doi.org/10.1007/BF02607055

Bihan, D. L., & Schild, T. (2017). Human brain MRI at 500 MHz, scientific perspectives and technological challenges. *Superconductor Science and Technology*, *30*(3), 033003. https://doi.org/10.1088/1361-6668/30/3/033003

Binder, J. R. (1997). Neuroanatomy of language processing studied with functional MRI. *Clinical Neuroscience (New York, N.Y.)*, *4*(2), 87–94.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex*, *19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *17*(1), 353–362.

Binder, J. R., McKiernan, K. A., Parsons, M. E., Westbury, C. F., Possing, E. T., Kaufman, J. N., & Buchanan, L. (2003). Neural Correlates of Lexical Access during Visual Word Recognition. *Journal of Cognitive Neuroscience*, *15*(3), 372–393. https://doi.org/10.1162/089892903321593108

Blanco-Elorrieta, E., Kastner, I., Emmorey, K., & Pylkkänen, L. (2018). Shared neural correlates for building phrases in signed and spoken language. *Scientific Reports*, *8*(1), Article 1. https://doi.org/10.1038/s41598-018-23915-0

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

Boccara, C. N., Nardin, M., Stella, F., O'Neill, J., & Csicsvari, J. (2019). The entorhinal cognitive map is attracted to goals. *Science*, *363*(6434), 1443–1447. https://doi.org/10.1126/science.aav4837

Boeckx, C. (2013). Merge: Biolinguistic Considerations. *English Linguistics*, *30*(2), 463–484. https://doi.org/10.9793/elsj.30.2_463

Bouillaud, J. (1825). Recherches cliniques propres à démontrer que la perte de la parole correspond à la lésion des lobules antérieures du cerveau, et à confirmer l'opinion de M. Gall sur le siège de l'organe du language articulé. *Archives Généralede Médecine*, *3*, 25–45.

Branch, C., Milner, B., & Rasmussen, T. (1964). Intracarotid Sodium Amytal for the Lateralization of Cerebral Speech Dominance: Observations in 123 Patients. *Journal of Neurosurgery*, *21*(5), 399–405. https://doi.org/10.3171/jns.1964.21.5.0399

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, *120*(2), 163–173. https://doi.org/10.1016/j.bandl.2010.04.002

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

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

Broca, P. (1861). Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bulletin et mémoires de la Société Anatomique de Paris*, *6*, 330–357.

Broca, P. (1865). Sur le siège de la faculté du langage articulé. *Bulletins et Mémoires de la Société d'Anthropologie de Paris*, *6*(1), 377–393. https://doi.org/10.3406/bmsap.1865.9495

Brooks, T. L., & Cid de Garcia, D. (2015). Evidence for morphological composition in compound words using MEG. *Frontiers in Human Neuroscience*, *9*. https://www.frontiersin.org/articles/10.3389/fnhum.2015.00215

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*, *41*(S6), 1318–1352. https://doi.org/10.1111/cogs.12461

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

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

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Univ of California Press.

Burroughs, A., Kazanina, N., & Houghton, C. (2021). Grammatical category and the neural processing of phrases. *Scientific Reports*, *11*(1), 2446. https://doi.org/10.1038/s41598-021-81901-5

Buzsáki, G., & Draguhn, A. (2004). Neuronal Oscillations in Cortical Networks. *Science*, *304*(5679), 1926–1929. https://doi.org/10.1126/science.1099745

Calvo Tapia, C., Tyukin, I., & Makarov, V. A. (2020). Universal principles justify the existence of concept cells. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-64466-7

Campadelli, P., Casiraghi, E., Ceruti, C., & Rozza, A. (2015). Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework. *Mathematical Problems in Engineering*, *2015*, e759567. https://doi.org/10.1155/2015/759567

Caramazza, A. (1997). How Many Levels of Processing Are There in Lexical Access? *Cognitive Neuropsychology*, *14*(1), 177–208. https://doi.org/10.1080/026432997381664

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), Article 1. https://doi.org/10.1038/nrn3136

Carpenter, P. A., Just, M. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1999). Time course of fMRI-activation in language and spatial networks during sentence comprehension. *Neuroimage*, *10*(2), 216–224. https://doi.org/10.1006/nimg.1999.0465

Carreira-Perpinán, M. A. (1997). *A Review of Dimension Reduction Techniques*.

Carroll, L. (1871). *Through the Looking-Glass and What Alice Found There*.

Caucheteux, C., Gramfort, A., & King, J.-R. (2021a). GPT-2's activations predict the degree of semantic comprehension in the human brain. *BioRxiv*, 2021.04.20.440622. https://doi.org/10.1101/2021.04.20.440622

Caucheteux, C., Gramfort, A., & King, J.-R. (2021b). Disentangling syntax and semantics in the brain with deep networks. *International Conference on Machine Learning*, 1336–1348. http://proceedings.mlr.press/v139/caucheteux21a.html

Caucheteux, C., & King, J.-R. (2020). *Language processing in brains and deep neural networks: Computational convergence and its limits* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.07.03.186288

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), Article 1. https://doi.org/10.1038/s42003-022-03036-1

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

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

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

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 103–111.

Chomsky, N. (1957). *Syntactic structures*. Mouton publishers. http://217.64.17.124:8080/xmlui/handle/123456789/557

Chomsky, N. (1965). *Aspects of the theory of syntax*.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, *3*(1), 1–15.

Chomsky, N. (1993). A minimalist program for linguistic theory. *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*.

Chomsky, N. (2013). Problems of projection. *Lingua*, *130*, 33–49. https://doi.org/10.1016/j.lingua.2012.12.003

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

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

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., & Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, *487*(7405), 51–56. https://doi.org/10.1038/nature11129

Churchland, P. S., & Sejnowski, T. J. (1988). Perspectives on cognitive neuroscience. *Science*, *242*(4879), 741–745.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What Does BERT Look at? An Analysis of BERT's Attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276–286.

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F., & Wattenberg, M. (2019). Visualizing and Measuring the Geometry of BERT. *ArXiv:1906.02715 [Cs, Stat]*. http://arxiv.org/abs/1906.02715

Cohen, D. (1968). Magnetoencephalography: Evidence of magnetic fields produced by alpha-rhythm currents. *Science (New York, N.Y.), 161*(3843), 784–786. https://doi.org/10.1126/science.161.3843.784

Cohen, M. S., & Bookheimer, S. Y. (1994). Localization of brain function using magnetic resonance imaging. *Trends in Neurosciences*, *17*(7), 268–277. https://doi.org/10.1016/0166-2236(94)90055-8

Coltheart, M. (2005). Modeling Reading: The Dual-Route Approach. In *The Science of Reading: A Handbook* (pp. 6–23). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470757642.ch1

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

Cressant, A., Muller, R. U., & Poucet, B. (2002). Remapping of place cell firing patterns after maze rotations. *Experimental Brain Research*, *143*(4), 470–479. https://doi.org/10.1007/s00221-002-1013-0

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314. https://doi.org/10.1007/BF02551274

Damasio, H., & Damasio, A. R. (1980). The anatomical basis of conduction aphasia. *Brain: A Journal of Neurology*, *103*(2), 337–350. https://doi.org/10.1093/brain/103.2.337

David, S. V., Vinje, W. E., & Gallant, J. L. (2004). Natural stimulus statistics alter the receptive field structure of v1 neurons. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *24*(31), 6991–7006. https://doi.org/10.1523/JNEUROSCI.1422-04.2004

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. *Neuron*, *88*(1), 2–19. https://doi.org/10.1016/j.neuron.2015.09.019

Del Giudice, M. (2021). Effective Dimensionality: A Tutorial. *Multivariate Behavioral Research*, *56*(3), 527–542. https://doi.org/10.1080/00273171.2020.1743631

Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition*, *135*, 103569. https://doi.org/10.1016/j.bandc.2019.05.007

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). The Representation of Semantic Information Across Human Cerebral Cortex During Listening Versus Reading Is Invariant to Stimulus Modality. *Journal of Neuroscience*, *39*(39), 7722–7736. https://doi.org/10.1523/JNEUROSCI.0675-19.2019

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

DeYoe, E. A., Bandettini, P., Neitz, J., Miller, D., & Winans, P. (1994). Functional magnetic resonance imaging (FMRI) of the human brain. *Journal of Neuroscience Methods*, *54*(2), 171–187. https://doi.org/10.1016/0165-0270(94)90191-0

Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Cortical Measures of Phoneme-Level Speech Encoding Correlate with the Perceived Clarity of Natural Speech. *ENeuro*, *5*(2), ENEURO.0084-18.2018. https://doi.org/10.1523/ENEURO.0084-18.2018

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), Article 1. https://doi.org/10.1038/nn.4186

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*. https://www.frontiersin.org/articles/10.3389/fnhum.2014.00311

Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron*, *109*(19), 3055–3068. https://doi.org/10.1016/j.neuron.2021.07.011

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

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

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Elmoznino, E., & Bonner, M. F. (2022). *High-performing neural network models of visual cortex benefit from high latent dimensionality* (p. 2022.07.13.499969). bioRxiv. https://doi.org/10.1101/2022.07.13.499969

Embick, D., & Poeppel, D. (2015). Towards a computational(ist) neurobiology of language: Correlational, integrated, and explanatory neurolinguistics. *Language, Cognition and Neuroscience*, *30*(4), 357–366. https://doi.org/10.1080/23273798.2014.980750

Facco, E., d'Errico, M., Rodriguez, A., & Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-11873-y

Fan, A., Grave, E., & Joulin, A. (2019). Reducing Transformer Depth on Demand with Structured Dropout. *International Conference on Learning Representations*.

Fedorenko, E., & Blank, I. A. (2020). Broca's Area Is Not a Natural Kind. *Trends in Cognitive Sciences*, *24*(4), 270–284. https://doi.org/10.1016/j.tics.2020.01.001

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

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

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

Fisher, R. A. (1936). Design of experiments. *British Medical Journal*, *1*(3923), 554.

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258-1270.e11. https://doi.org/10.1016/j.neuron.2022.01.005

Flick, G., Oseki, Y., Kaczmarek, A. R., Al Kaabi, M., Marantz, A., & Pylkkänen, L. (2018). Building words and phrases in the left temporal lobe. *Cortex*, *106*, 213–236. https://doi.org/10.1016/j.cortex.2018.06.004

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Frank, S. L., & Yang, J. (2018). Lexical representation explains cortical entrainment during speech comprehension. *PLOS ONE*, *13*(5), e0197304. https://doi.org/10.1371/journal.pone.0197304

Frankland, S. M., & Greene, J. D. (2015). An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences*, *112*(37), 11732–11737. https://doi.org/10.1073/pnas.1421236112

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

Frankland, S. M., & Greene, J. D. (2020b). Two Ways to Build a Thought: Distinct Forms of Compositional Semantic Representation across Brain Regions. *Cerebral Cortex*, *30*(6), 3838–3855. https://doi.org/10.1093/cercor/bhaa001

Frazier, L., & Clifton, C. (1996). *Construal*. Mit Press.

Frenzel, S., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2011). Conflicts in language processing: A new perspective on the N400–P600 distinction. *Neuropsychologia*, *49*(3), 574–579. https://doi.org/10.1016/j.neuropsychologia.2010.12.003

Friederici, A. D. (2011). The Brain Basis of Language Processing: From Structure to Function. *Physiological Reviews*, *91*(4), 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Friederici, A. D. (2012). The cortical language circuit: From auditory perception to sentence comprehension. *Trends in Cognitive Sciences*, *16*(5), 262–268. https://doi.org/10.1016/j.tics.2012.04.001

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

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

Friederici, A. D., Meyer, M., & von Cramon, D. Y. (2000). Auditory language comprehension: An event-related fMRI study on the processing of syntactic and lexical information. *Brain and Language*, *75*(3), 289–300.

Friederici, A. D., Opitz, B., & von Cramon, D. Y. (2000). Segregating Semantic and Syntactic Aspects of Processing in the Human Brain: An fMRI Investigation of Different Word Types. *Cerebral Cortex*, *10*(7), 698–705. https://doi.org/10.1093/cercor/10.7.698

Fukui, N. (2017). *Merge in the Mind-Brain: Essays on Theoretical Linguistics and the Neuroscience of Language*. Taylor & Francis.

Fyhn, M., Hafting, T., Treves, A., Moser, M.-B., & Moser, E. I. (2007). Hippocampal remapping and grid realignment in entorhinal cortex. *Nature*, *446*(7132), Article 7132. https://doi.org/10.1038/nature05601

Fyshe, A. (2020). Studying language in context using the temporal generalization method. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791). https://doi.org/10.1098/rstb.2018.0531

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

Gallego, J. A., Perich, M. G., Miller, L. E., & Solla, S. A. (2017). Neural Manifolds for the Control of Movement. *Neuron*, *94*(5), 978–984. https://doi.org/10.1016/j.neuron.2017.05.025

Gao, P., Trautmann, E., Yu, B., Santhanam, G., Ryu, S., Shenoy, K., & Ganguli, S. (2017). A theory of multineuronal dimensionality, dynamics and measurement. *BioRxiv*, 214262. https://doi.org/10.1101/214262

Gayler, R. W. (2004). Vector Symbolic Architectures answer Jackendoff's challenges for cognitive neuroscience. *ArXiv:Cs/0412059*. http://arxiv.org/abs/cs/0412059

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, *233*(4771), 1416–1419.

Geschwind, N. (1965). Disconnexion syndromes in animals and man. I. *Brain: A Journal of Neurology*, *88*(2), 237–294.

Geschwind, N. (1970). The Organization of Language and the Brain: Language disorders after brain damage help in elucidating the neural basis of verbal behavior. *Science*, *170*(3961), 940–944.

Goldberg, Y. (2019). Assessing BERT's Syntactic Abilities. *ArXiv:1901.05287 [Cs]*. http://arxiv.org/abs/1901.05287

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

Goldstein, A., Dabush, A., Aubrey, B., Schain, M., Nastase, S. A., Zada, Z., Ham, E., Hong, Z., Feder, A., Gazula, H., Buchnik, E., Doyle, W., Devore, S., Dugan, P., Friedman, D., Brenner, M., Hassidim, A., Devinsky, O., Flinker, A., & Hasson, U. (2022a). Brain embeddings with shared geometry to artificial contextual embeddings, as a code for representing language in the human brain. *BioRxiv*. https://doi.org/10.1101/2022.03.01.482586

Goldstein, A., Dabush, A., Aubrey, B., Schain, M., Nastase, S. A., Zada, Z., Ham, E., Hong, Z., Feder, A., Gazula, H., Buchnik, E., Doyle, W., Devore, S., Dugan, P., Friedman, D., Brenner, M., Hassidim, A., Devinsky, O., Flinker, A., & Hasson, U. (2022b). *Brain embeddings with shared geometry to artificial contextual embeddings, as a code for representing language in the human brain* [Preprint]. Neuroscience. https://doi.org/10.1101/2022.03.01.482586

Goldstein, A., Ham, E., Nastase, S. A., Zada, Z., Grinstein-Dabus, A., Aubrey, B., Schain, M., Gazula, H., Feder, A., Doyle, W., Devore, S., Dugan, P., Friedman, D., Brenner, M., Hassidim, A., Devinsky, O., Flinker, A., Levy, O., & Hasson, U. (2022). Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. *BioRxiv*. https://doi.org/10.1101/2022.07.11.499562

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), Article 3. https://doi.org/10.1038/s41593-022-01026-4

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, *27*. https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

Gorban, A. N., Makarov, V. A., & Tyukin, I. Y. (2019). The unreasonable effectiveness of small neural ensembles in high-dimensional brain. *Physics of Life Reviews*, *29*, 55–88. https://doi.org/10.1016/j.plrev.2018.09.005

Gorban, A. N., Makarov, V. A., & Tyukin, I. Y. (2020). High-Dimensional Brain in a High-Dimensional World: Blessing of Dimensionality. *Entropy*, *22*(1), Article 1. https://doi.org/10.3390/e22010082

Goucha, T., & Friederici, A. D. (2015). The language skeleton after dissecting meaning: A functional segregation within Broca's Area. *NeuroImage*, *114*, 294–302. https://doi.org/10.1016/j.neuroimage.2015.04.011

GPT-3. (2020). A robot wrote this entire article. Are you scared yet, human? *The Guardian*.

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

Granata, D., & Carnevale, V. (2016). Accurate Estimation of the Intrinsic Dimension Using Graph Distances: Unraveling the Geometric Complexity of Datasets. *Scientific Reports*, *6*(1), Article 1. https://doi.org/10.1038/srep31377

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, 1–13. https://doi.org/10.1038/s41562-022-01316-8

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., & others. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471–476.

Graves, R. E. (1997). The Legacy of the Wernicke-Lichtheim Model. *Journal of the History of the Neurosciences*, *6*(1), 3–20. https://doi.org/10.1080/09647049709525682

Greibach, S. A. (1978). Comments on universal and left universal grammars, context-sensitive languages, and context-free grammar forms. *Information and Control*, *39*(2), 135–142. https://doi.org/10.1016/S0019-9958(78)90799-4

Guerguiev, J., Lillicrap, T. P., & Richards, B. A. (2017). Towards deep learning with segregated dendrites. *ELife*, *6*, e22901. https://doi.org/10.7554/eLife.22901

Guillem, Fran., N'kaoua, B., Rougier, A., & Claverie, B. (1995). Intracranial topography of event-related potentials (N400/P600) elicited during a continuous recognition memory task. *Psychophysiology*, *32*(4), 382–392. https://doi.org/10.1111/j.1469-8986.1995.tb01221.x

Gwilliams, L. (2020). How the brain composes morphemes into meaning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190311. https://doi.org/10.1098/rstb.2019.0311

Gwilliams, L., & King, J.-R. (2020). Recurrent processes support a cascade of hierarchical decisions. *ELife*, *9*, e56603. https://doi.org/10.7554/eLife.56603

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

Hagoort, P. (2019). The neurobiology of language beyond single-word processing. *Science*, *366*(6461), 55–58. https://doi.org/10.1126/science.aax0289

Hahn, M., & Baroni, M. (2019). Tabula Nearly Rasa: Probing the Linguistic Knowledge of Character-level Neural Language Models Trained on Unsegmented Text. *Transactions of the Association for Computational Linguistics*, *7*, 467–484. https://doi.org/10.1162/tacl_a_00283

Hahne, A., & Jescheniak, J. D. (2001a). What's left if the Jabberwock gets the semantics? An ERP investigation into semantic and syntactic processes during auditory sentence comprehension. *Brain Research. Cognitive Brain Research*, *11*(2), 199–212. https://doi.org/10.1016/s0926-6410(00)00071-9

Hahne, A., & Jescheniak, J. D. (2001b). What's left if the Jabberwock gets the semantics? An ERP investigation into semantic and syntactic processes during auditory sentence comprehension. *Brain Research. Cognitive Brain Research*, *11*(2), 199–212. https://doi.org/10.1016/s0926-6410(00)00071-9

Hale, J. T., Campanelli, L., Li, J., Bhattasali, S., Pallier, C., & Brennan, J. R. (2021). Neuro-computational models of language processing. *Annual Review of Linguistics*. https://doi.org/10.1146/lingbuzz/006147

Hale, J. T., Campanelli, L., Li, J., Bhattasali, S., Pallier, C., & Brennan, J. R. (2022). Neurocomputational Models of Language Processing. *Annual Review of Linguistics*, *8*(1), 427–446. https://doi.org/10.1146/annurev-linguistics-051421-020803

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573–582. https://doi.org/10.1080/23273798.2018.1499946

Harris, Z. S. (Zellig S. (1951). *Methods in structural linguistics*. Chicago : University of Chicago Press. http://archive.org/details/methodsinstructu0000harr

Hart, J., Kraut, M. A., Kremen, S., Soher, B., & Gordon, B. (2000). Neural substrates of orthographic lexical access as demonstrated by functional brain imaging. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, *13*(1), 1–7.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, *105*(3), 416–434. https://doi.org/10.1016/j.neuron.2019.12.002

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, *298*(5598), 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Haxby, J. V., Gobbini, M. I., & Nastase, S. A. (2020). Naturalistic stimuli reveal a dominant role for agentic action in visual representation. *NeuroImage*, *216*, 116561. https://doi.org/10.1016/j.neuroimage.2020.116561

He, Y., Sommer, J., Hansen-Schirra, S., & Nagels, A. (2022). *Negation impacts sentence processing in the N400 and later time windows: Evidence from multivariate pattern analysis of EEG*. PsyArXiv. https://doi.org/10.31234/osf.io/8rbw3

Heavey, C. L., Moynihan, S. A., Brouwers, V. P., Lapping-Carr, L., Krumm, A. E., Kelsey, J. M., Turner, D. K., & Hurlburt, R. T. (2019). Measuring the frequency of inner-experience characteristics by self-report: The Nevada Inner Experience Questionnaire. *Frontiers in Psychology*, *9*, 2615.

Hebb, D. (1949). *The organization of behavior; a neuropsychological theory.*

Heer, W. A. de, Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The Hierarchical Cortical Organization of Human Speech Processing. *Journal of Neuroscience*, *37*(27), 6539–6557. https://doi.org/10.1523/JNEUROSCI.3267-16.2017

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. https://doi.org/10.1073/pnas.2201968119

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), Article 5. https://doi.org/10.1038/nrn2113

Hinton, G. E. (1984). *Distributed representations*.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv Preprint ArXiv:1207.0580*.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holleman, G. A., Hooge, I. T. C., Kemner, C., & Hessels, R. S. (2020). The 'Real-World Approach' and Its Problems: A Critique of the Term Ecological Validity. *Frontiers in Psychology*, *11*, 721. https://doi.org/10.3389/fpsyg.2020.00721

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

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257. https://doi.org/10.1016/0893-6080(91)90009-T

Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2022). Meta-Learning in Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5149–5169. https://doi.org/10.1109/TPAMI.2021.3079209

Hounsfield, G. N. (1980). Computed medical imaging. *Science*, *210*(4465), 22–28.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106.

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2006). Syntactic and Semantic Modulation of Neural Activity during Auditory Sentence Comprehension. *Journal of Cognitive Neuroscience*, *18*(4), 665–679. https://doi.org/10.1162/jocn.2006.18.4.665

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

Humphries, C., Willard, K., Buchsbaum, B., & Hickok, G. (2001). Role of anterior temporal cortex in auditory sentence comprehension: An fMRI study. *Neuroreport*, *12*, 1749–1752. https://doi.org/10.1097/00001756-200106130-00046

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), Article 7600. https://doi.org/10.1038/nature17637

Huth, A. G., Lee, T., Nishimoto, S., Bilenko, N. Y., Vu, A. T., & Gallant, J. L. (2016). Decoding the Semantic Content of Natural Movies from Human Brain Activity. *Frontiers in Systems Neuroscience*, *10*. https://doi.org/10.3389/fnsys.2016.00081

Illing, B., Ventura, J., Bellec, G., & Gerstner, W. (2021). Local plasticity rules can learn deep representations using self-supervised contrastive predictions. *Advances in Neural Information Processing Systems*, *34*, 30365–30379.

https://proceedings.neurips.cc/paper/2021/hash/feade1d2047977cd0cefdafc40175a99-Abstract.html

Isitan, C., Yan, Q., Spencer, D. D., & Alkawadri, R. (2020). Brief history of electrical cortical stimulation: A journey in time from Volta to Penfield. *Epilepsy Research*, *166*, 106363. https://doi.org/10.1016/j.eplepsyres.2020.106363

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

Jackendoff, R. S. (1972). *Semantic interpretation in generative grammar*.

Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castañeda, A. G., Beattie, C., Rabinowitz, N. C., Morcos, A. S., Ruderman, A., Sonnerat, N., Green, T., Deason, L., Leibo, J. Z., Silver, D., Hassabis, D., Kavukcuoglu, K., & Graepel, T. (2019). Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, *364*(6443), 859–865. https://doi.org/10.1126/science.aau6249

Jain, S., & Huth, A. (2018). Incorporating Context into Language Encoding Models for fMRI. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper/2018/hash/f471223d1a1614b58a7dc45c9d01df19-Abstract.html

Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. https://doi.org/10.18653/v1/P19-1356

Jefferies, E. (2013). The neural basis of semantic cognition: Converging evidence from neuropsychology, neuroimaging and TMS. *Cortex*, *49*(3), 611–625. https://doi.org/10.1016/j.cortex.2012.10.008

Jobard, G., Crivello, F., & Tzourio-Mazoyer, N. (2003). Evaluation of the dual route theory of reading: A metanalysis of 35 neuroimaging studies. *NeuroImage*, *20*(2), 693–712. https://doi.org/10.1016/S1053-8119(03)00343-4

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science*, *274*(5284), 114–116.

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, *15*(2), 159–201. https://doi.org/10.1080/016909600386084

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

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *Journal of Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory

Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630-644.e16. https://doi.org/10.1016/j.neuron.2018.03.044

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

King, J.-R., Charton, F., Lopez-Paz, D., & Oquab, M. (2020). Back-to-back regression: Disentangling the influence of correlated factors from multivariate observations. *NeuroImage*, *220*, 117028. https://doi.org/10.1016/j.neuroimage.2020.117028

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

Kingma, D. P., & Welling, M. (2014). *Auto-Encoding Variational Bayes* (arXiv:1312.6114; Version 10). arXiv. https://doi.org/10.48550/arXiv.1312.6114

Kleyko, D., Rachkovskij, D. A., Osipov, E., & Rahimi, A. (2022). A Survey on Hyperdimensional Computing aka Vector Symbolic Architectures, Part I: Models and Data Transformations. *ACM Computing Surveys*. https://doi.org/10.1145/3538531

Kuperman, V., Dambacher, M., Nuthmann, A., & Kliegl, R. (2010). The effect of word position on eye-movements in sentence and paragraph reading. *Quarterly Journal of Experimental Psychology*, *63*(9), 1838–1857. https://doi.org/10.1080/17470211003602412

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity. *Science*, *207*(4427), 203–205. https://doi.org/10.1126/science.7350657

Laje, R., & Buonomano, D. V. (2013). Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nature Neuroscience*, *16*(7), 925–933. https://doi.org/10.1038/nn.3405

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Lakretz, Y., Desbordes, T., Hupkes, D., & Dehaene, S. (2021). *Causal Transformers Perform Below Chance on Recursive Nested Constructions, Unlike Humans* (arXiv:2110.07240). arXiv. http://arxiv.org/abs/2110.07240

Lakretz, Y., Desbordes, T., Hupkes, D., & Dehaene, S. (2022). Can Transformers Process Recursive Nested Constructions, Like Humans? *Proceedings of the 29th International Conference on Computational Linguistics*, 3226–3232. https://aclanthology.org/2022.coling-1.285

Lakretz, Y., Desbordes, T., King, J.-R., Crabbé, B., Oquab, M., & Dehaene, S. (2021a). Can RNNs learn Recursive Nested Subject-Verb Agreements? *ArXiv:2101.02258 [Cs]*. http://arxiv.org/abs/2101.02258

Lakretz, Y., Desbordes, T., King, J.-R., Crabbé, B., Oquab, M., & Dehaene, S. (2021b). Can RNNs learn Recursive Nested Subject-Verb Agreements? *ArXiv:2101.02258 [Cs]*. http://arxiv.org/abs/2101.02258

Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., & Dehaene, S. (2020). Exploring Processing of Nested Dependencies in Neural-Network Language Models and Humans. *ArXiv:2006.11098 [Cs]*. http://arxiv.org/abs/2006.11098

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019). The emergence of number and syntax units in LSTM language models. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 11–20. https://doi.org/10.18653/v1/N19-1002

Landa, B., Zhang, T. T. C. K., & Kluger, Y. (2021). *Biwhitening Reveals the Rank of a Count Matrix* (arXiv:2103.13840). arXiv. https://doi.org/10.48550/arXiv.2103.13840

Langston, R. F., Ainge, J. A., Couey, J. J., Canto, C. B., Bjerknes, T. L., Witter, M. P., Moser, E. I., & Moser, M.-B. (2010). Development of the Spatial Representation System in the Rat. *Science*, *328*(5985), 1576–1580. https://doi.org/10.1126/science.1188210

Lasnik, H., & Lidz, J. (2016). The Argument from the Poverty of the Stimulus. In I. Roberts (Ed.), *The Oxford Handbook of Universal Grammar* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199573776.013.10

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), Article 12. https://doi.org/10.1038/nrn2532

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Legate, J. A., & Yang, C. D. (2002). Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, *19*(1–2), 151–162. https://doi.org/10.1515/tlir.19.1-2.151

Lenneberg, E. H. (1967). The biological foundations of language. *Hospital Practice*, *2*(12), 59–67.

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

Lewis, A. G., Wang, L., & Bastiaansen, M. (2015). Fast oscillatory dynamics during language comprehension: Unification versus maintenance and prediction? *Brain and Language*, *148*, 51–63. https://doi.org/10.1016/j.bandl.2015.01.003

Lewis, R. L., & Vasishth, S. (2005). An Activation-Based Model of Sentence Processing as Skilled Memory Retrieval. In *Cognitive Science*. Routledge.

Li, H. (2022). Language models: Past, present, and future. *Communications of the ACM*, *65*(7), 56–63. https://doi.org/10.1145/3490443

Lichtheim, L. (1885). On aphasia. *Brain, 7*, 433–484.

Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications, 7*, 13276. https://doi.org/10.1038/ncomms13276

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

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

Lo, C.-W., Tung, T.-Y., Ke, A. H., & Brennan, J. R. (2022). Hierarchy, Not Lexical Regularity, Modulates Low-Frequency Neural Synchrony During Language Comprehension. *Neurobiology of Language, 3*(4), 538–555. https://doi.org/10.1162/nol_a_00077

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature, 453*(7197), Article 7197. https://doi.org/10.1038/nature06976

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

Maass, W., Natschläger, T., & Markram, H. (2002). Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations. *Neural Computation, 14*(11), 2531–2560. https://doi.org/10.1162/089976602760407955

Macdonald, R. G., & Tatler, B. W. (2013). Do as eye say: Gaze cueing and language in a real-world social interaction. *Journal of Vision*, *13*(4), 6. https://doi.org/10.1167/13.4.6

Machens, C. K., Romo, R., & Brody, C. D. (2010). Functional, But Not Anatomical, Separation of "What" and "When" in Prefrontal Cortex. *Journal of Neuroscience*, *30*(1), 350–360. https://doi.org/10.1523/JNEUROSCI.3276-09.2010

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), Article 7474. https://doi.org/10.1038/nature12742

Manuel, F. (1979). No Man Alone: A Neurosurgeon's Life by Wilder Penfield. *The Canadian Historical Review*, *60*(4), 510–511.

Marcus, G. (2020). *The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence* (arXiv:2002.06177). arXiv. https://doi.org/10.48550/arXiv.2002.06177

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

Marr, D., & Poggio, T. (1976). *From understanding computation to understanding neural circuitry*.

Marshall, J. C., & Newcombe, F. (1973). Patterns of paralexia: A psycholinguistic approach. *Journal of Psycholinguistic Research*, *2*(3), 175–199. https://doi.org/10.1007/BF01067101

Martin, A. E. (2020). A Compositional Neural Architecture for Language. *Journal of Cognitive Neuroscience*, *32*(8), 1407–1427. https://doi.org/10.1162/jocn_a_01552

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

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., & Sagot, B. (2020). CamemBERT: A Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203–7219. https://doi.org/10.18653/v1/2020.acl-main.645

Matchin, W., & Hickok, G. (2020). The Cortical Organization of Syntax. *Cerebral Cortex*, *30*(3), 1481–1498. https://doi.org/10.1093/cercor/bhz180

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

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(42), 25966–25974. https://doi.org/10.1073/pnas.1910416117

McCoy, R. T., Linzen, T., Dunbar, E., & Smolensky, P. (2018, September 27). *RNNs implicitly implement tensor-product representations*. International Conference on Learning Representations. https://openreview.net/forum?id=BJx0sjC5FX

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

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

McMahon, D. B. T., Russ, B. E., Elnaiem, H. D., Kurnikova, A. I., & Leopold, D. A. (2015). Single-Unit Activity during Natural Vision: Diversity, Consistency, and Spatial Sensitivity among AF Face Patch Neurons. *Journal of Neuroscience*, *35*(14), 5537–5548. https://doi.org/10.1523/JNEUROSCI.3825-14.2015

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). Pointer Sentinel Mixture Models. *ArXiv:1609.07843 [Cs]*. http://arxiv.org/abs/1609.07843

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the 2013 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, 746–751. https://www.aclweb.org/anthology/N13-1090

Miller, G. (2008). Growing Pains for fMRI. *Science*, *320*(5882), 1412–1414. https://doi.org/10.1126/science.320.5882.1412

Miller, K. J., Sorensen, L. B., Ojemann, J. G., & Nijs, M. den. (2009). Power-Law Scaling in the Brain Surface Electric Potential. *PLOS Computational Biology*, *5*(12), e1000609. https://doi.org/10.1371/journal.pcbi.1000609

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., & King, J.-R. (2022). *Toward a realistic model of speech processing in the brain with self-supervised learning* (arXiv:2206.01685). arXiv. http://arxiv.org/abs/2206.01685

Mirman, D., Chen, Q., Zhang, Y., Wang, Z., Faseyitan, O. K., Coslett, H. B., & Schwartz, M. F. (2015). Neural organization of spoken language revealed by lesion-symptom mapping. *Nature Communications*, *6*, 6762. https://doi.org/10.1038/ncomms7762

Mitchell, J., & Lapata, M. (2008). Vector-based Models of Semantic Composition. *Proceedings of ACL-08: HLT*, 236–244. https://aclanthology.org/P08-1028

Mitchell, T. M. (1980). *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research ….

Molinaro, N., Vespignani, F., & Job, R. (2008). A deeper reanalysis of a superficial feature: An ERP study on agreement violations. *Brain Research*, *1228*, 161–176. https://doi.org/10.1016/j.brainres.2008.06.064

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

Moonen, C. T. W., van Zijl, P. C. M., Frank, J. A., Le Bihan, D., & Becker, E. D. (1990). Functional Magnetic Resonance Imaging in Medicine and Physiology. *Science*, *250*(4977), 53–61. https://doi.org/10.1126/science.2218514

Moser, E. I., Kropff, E., Moser, M.-B., & others. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, *31*(1), 69–89.

Musk, E. & others. (2019). An integrated brain-machine interface platform with thousands of channels. *Journal of Medical Internet Research*, *21*(10), e16194.

Nasios, G., Dardiotis, E., & Messinis, L. (2019). From Broca and Wernicke to the Neuromodulation Era: Insights of Brain Language Networks for Neurorehabilitation. *Behavioural Neurology*, *2019*, 9894571. https://doi.org/10.1155/2019/9894571

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., Nguyen, M., Chang, C. H. C., Baldassano, C., Lositsky, O., Simony, E., Chow, M. A., Leong, Y. C., Brooks, P. P., Micciche, E., … Hasson, U. (2020). Narratives: FMRI data for evaluating models of naturalistic language comprehension. *BioRxiv*, 2020.12.23.424091. https://doi.org/10.1101/2020.12.23.424091

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

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

Neufeld, C., Kramer, S. E., Lapinskaya, N., Heffner, C. C., Malko, A., & Lau, E. F. (2016). The Electrophysiology of Basic Phrase Building. *PLOS ONE*, *11*(10), e0158446. https://doi.org/10.1371/journal.pone.0158446

Newell, A., & Simon, H. A. (1961). *GPS, a program that simulates human thought*. RAND CORP SANTA MONICA CALIF.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., & Srebro, N. (2018). Towards understanding the role of over-parametrization in generalization of neural networks. *ArXiv Preprint ArXiv:1805.12076*.

Nobre, A. C., Allison, T., & McCarthy, G. (1994). Word recognition in the human inferior temporal lobe. *Nature*, *372*(6503), Article 6503. https://doi.org/10.1038/372260a0

Nozaradan, S., Peretz, I., & Mouraux, A. (2012). Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(49), 17572–17581. https://doi.org/10.1523/JNEUROSCI.3203-12.2012

Obleser, J., & Kayser, C. (2019). Neural Entrainment and Attentional Selection in the Listening Brain. *Trends in Cognitive Sciences*, *23*(11), 913–926. https://doi.org/10.1016/j.tics.2019.08.004

Ogawa, S., Lee, T.-M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, *87*(24), 9868–9872.

Oh, B.-D., Clark, C., & Schuler, W. (2022). Comparison of Structural Parsers and Neural Language Models as Surprisal Estimators. *Frontiers in Artificial Intelligence*, *5*, 777963. https://doi.org/10.3389/frai.2022.777963

Ojemann, G., Ojemann, J., Lettich, E., & Berger, M. (1989). Cortical language localization in left, dominant hemisphere: An electrical stimulation mapping investigation in 117 patients. *Journal of Neurosurgery*, *71*(3), 316–326.

Osterhout, L. E. (1991). *Event-related brain potentials elicited during sentence comprehension*. 1.

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 604–624. https://doi.org/10.1109/TNNLS.2020.2979670

Pagin, P., & Westerståhl, D. (2010). Compositionality I: Definitions and Variants. *Philosophy Compass*, *5*(3), 250–264. https://doi.org/10.1111/j.1747-9991.2009.00228.x

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

Pasquiou, A., Lakretz, Y., Hale, J., Thirion, B., & Pallier, C. (2022). *Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps* (arXiv:2207.03380). arXiv. https://doi.org/10.48550/arXiv.2207.03380

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library* (arXiv:1912.01703). arXiv. https://doi.org/10.48550/arXiv.1912.01703

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, *95*(1), e41–e74. https://doi.org/10.1353/lan.2019.0009

Pattamadilok, C., Dehaene, S., & Pallier, C. (2016). A role for left inferior frontal and posterior superior temporal cortex in extracting a syntactic tree from a sentence. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *75*, 44–55. https://doi.org/10.1016/j.cortex.2015.11.012

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

Penfield, W. (1959). *Speech and brain-mechanisms*. Princeton, N.J. : Princeton University Press. http://archive.org/details/speechbrainmecha0000penf

Penfield, W., & Rasmussen, T. (1950). *The cerebral cortex of man; a clinical study of localization of function.*

Piattelli-Palmarini, M. (1980). *Language and learning: The debate between Jean Piaget and Noam Chomsky*.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1), 73–193. https://doi.org/10.1016/0010-0277(88)90032-7

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

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*(3), 623–641. https://doi.org/10.1109/72.377968

Poeppel, D., Emmorey, K., Hickok, G., & Pylkkänen, L. (2012). Towards a New Neurobiology of Language. *Journal of Neuroscience*, *32*(41), 14125–14131. https://doi.org/10.1523/JNEUROSCI.3244-12.2012

Poeppel, D., & Hickok, G. (2004). Towards a new functional anatomy of language. *Cognition*, *92*(1–2), 1–12. https://doi.org/10.1016/j.cognition.2003.11.001

Poggio, T., Torre, V., & Koch, C. (1987). Computational vision and regularization theory. In M. A. Fischler & O. Firschein (Eds.), *Readings in Computer Vision* (pp. 638–643). Morgan Kaufmann. https://doi.org/10.1016/B978-0-08-051581-6.50061-1

Poldrack, R. A., Wagner, A. D., Prull, M. W., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. (1999). Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *NeuroImage*, *10*(1), 15–35. https://doi.org/10.1006/nimg.1999.0441

Popham, S., Boebinger, D., Ellis, D. P. W., Kawahara, H., & McDermott, J. H. (2018). Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nature Communications*, *9*(1), 2122. https://doi.org/10.1038/s41467-018-04551-8

Price, A. R., Bonner, M. F., Peelle, J. E., & Grossman, M. (2015). Converging Evidence for the Neuroanatomic Basis of Combinatorial Semantics in the Angular Gyrus. *Journal of Neuroscience*, *35*(7), 3276–3284.

Price, A. R., Peelle, J. E., Bonner, M. F., Grossman, M., & Hamilton, R. H. (2016). Causal Evidence for a Mechanism of Semantic Integration in the Angular Gyrus as Revealed by High-Definition Transcranial Direct Current Stimulation. *Journal of Neuroscience*, *36*(13), 3829–3838.

Price, C. J. (2000). The anatomy of language: Contributions from functional neuroimaging. *The Journal of Anatomy*, *197*(3), 335–359. https://doi.org/10.1046/j.1469-7580.2000.19730335.x

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

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

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

Pylkkänen, L., Bemis, D. K., & Blanco Elorrieta, E. (2014). Building phrases in language production: An MEG study of simple composition. *Cognition*, *133*(2), 371–384. https://doi.org/10.1016/j.cognition.2014.07.001

Quentin, R., King, J.-R., Sallard, E., Fishman, N., Thompson, R., Buch, E. R., & Cohen, L. G. (2019). Differential Brain Mechanisms of Selection and Maintenance of Information during Working Memory. *Journal of Neuroscience*, *39*(19), 3728–3740. https://doi.org/10.1523/JNEUROSCI.2764-18.2019

Quine, W. V. O. (1960). *Word and Object*. MIT Press.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), Article 7045. https://doi.org/10.1038/nature03687

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), Article 9. https://doi.org/10.1038/s41562-018-0406-4

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., & others. (2018). *Improving language understanding by generative pre-training*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & others. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* (arXiv:1910.10683). arXiv. http://arxiv.org/abs/1910.10683

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

Ray, S., & Maunsell, J. H. R. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biology*, *9*(4), e1000610. https://doi.org/10.1371/journal.pbio.1000610

Recanatesi, S., Farrell, M., Lajoie, G., Deneve, S., Rigotti, M., & Shea-Brown, E. (2021). Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-21696-1

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

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

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

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D.,

Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), Article 11. https://doi.org/10.1038/s41593-019-0520-2

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633. https://doi.org/10.1145/365628.365657

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*. California Univ San Diego La Jolla Inst for Cognitive Science.

Rumelhart, D. E., & McClelland, J. L. (1986). *On learning the past tenses of English verbs*.

Rumsey, J. M., Horwitz, B., Donohue, B. C., Nace, K., Maisog, J. M., & Andreason, P. (1997). Phonological and orthographic components of word recognition. A PET-rCBF study. *Brain*, *120*(5), 739–759. https://doi.org/10.1093/brain/120.5.739

Russin, J., Jo, J., O'Reilly, R. C., & Bengio, Y. (2019). Compositional generalization in a deep seq2seq model by separating syntax and semantics. *ArXiv:1904.09708 [Cs, Stat]*. http://arxiv.org/abs/1904.09708

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., & others. (2022). Photorealistic Text-to-

Image Diffusion Models with Deep Language Understanding. *ArXiv Preprint ArXiv:2205.11487*.

Sanchez-Giraldo, L. G., Laskar, M. N. U., & Schwartz, O. (2019). Normalization and pooling in hierarchical models of natural images. *Current Opinion in Neurobiology*, *55*, 65–72. https://doi.org/10.1016/j.conb.2019.01.008

Sato, S., & Smith, P. D. (1985). Magnetoencephalography. *Journal of Clinical Neurophysiology*, *2*(2), 173–192.

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.-S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., & Weiller, C. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(46), 18035–18040. https://doi.org/10.1073/pnas.0805234105

Savitch, W. J. (1987). Context-Sensitive Grammar and Natural Language Syntax. In W. J. Savitch, E. Bach, W. Marsh, & G. Safran-Naveh (Eds.), *The Formal Complexity of Natural Language* (pp. 358–368). Springer Netherlands. https://doi.org/10.1007/978-94-009-3401-6_15

Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., & Cox, D. D. (2019). On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, *2019*(12), 124020. https://doi.org/10.1088/1742-5468/ab3985

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

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *BioRxiv*, 407007. https://doi.org/10.1101/407007

Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, S089662732030605X. https://doi.org/10.1016/j.neuron.2020.07.040

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568. https://doi.org/10.1037/0033-295X.96.4.523

Senel, L. K., Utlu, I., Yucesoy, V., Koc, A., & Cukur, T. (2018). Semantic Structure and Interpretability of Word Embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1769–1779. https://doi.org/10.1109/TASLP.2018.2837384

Seo, D., Neely, R. M., Shen, K., Singhal, U., Alon, E., Rabaey, J. M., Carmena, J. M., & Maharbiz, M. M. (2016). Wireless recording in the peripheral nervous system with ultrasonic neural dust. *Neuron*, *91*(3), 529–539.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). FMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

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

Sharkey, N. E. (1992). *Connectionist Natural Language Processing: Readings from Connection Science*. Intellect Books.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), Article 7676. https://doi.org/10.1038/nature24270

Singer, Y., Teramoto, Y., Willmore, B. D., Schnupp, J. W., King, A. J., & Harper, N. S. (2018). Sensory cortex is optimized for prediction of future input. *ELife*, *7*, e31557. https://doi.org/10.7554/eLife.31557

Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., Zhang, E., Child, R., Aminabadi, R.

Y., Bernauer, J., Song, X., Shoeybi, M., He, Y., Houston, M., Tiwary, S., & Catanzaro, B. (2022). *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model* (arXiv:2201.11990). arXiv. https://doi.org/10.48550/arXiv.2201.11990

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

Smolensky, P., McCoy, R. T., Fernandez, R., Goldrick, M., & Gao, J. (2022). Neurocompositional computing: From the Central Paradox of Cognition to a new generation of AI systems. *ArXiv:2205.01128 [Cs]*. http://arxiv.org/abs/2205.01128

Sorscher, B., Ganguli, S., & Sompolinsky, H. (2021). The Geometry of Concept Learning. *BioRxiv*, 2021.03.21.436284. https://doi.org/10.1101/2021.03.21.436284

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

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

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2022). *Beyond the Imitation Game: Quantifying and extrapolating the*

capabilities of language models (arXiv:2206.04615). arXiv. http://arxiv.org/abs/2206.04615

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Steedman, M. (2001). *The syntactic process*. MIT press.

Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., Chen, S., Colonell, J., Gardner, R. J., Karsh, B., Kloosterman, F., Kostadinov, D., Mora-Lopez, C., O'Callaghan, J., Park, J., … Harris, T. D. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, *372*(6539), eabf4588. https://doi.org/10.1126/science.abf4588

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

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

Storks, S., Gao, Q., & Chai, J. Y. (2019). Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *ArXiv Preprint ArXiv:1904.01172*.

Stowe, L. A., Broere, C. A. J., Paans, A. M. J., Wijers, A. A., Mulder, G., Vaalburg, W., & Zwarts, F. (1998). Localizing components of a complex task: Sentence processing and working memory. *NeuroReport, 9*(13), 2995–2999.

Stowe, L. A., Kaan, E., Sabourin, L., & Taylor, R. C. (2018). The sentence wrap-up dogma. *Cognition, 176*, 232–247. https://doi.org/10.1016/j.cognition.2018.03.011

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*(2), 135.

Szostak, K. M., Grand, L., & Constandinou, T. G. (2017). Neural Interfaces for Intracortical Recording: Requirements, Fabrication Methods, and Characteristics. *Frontiers in Neuroscience*, *11*, 665. https://doi.org/10.3389/fnins.2017.00665

Talairach, J., Bancaud, J., Bonis, A., Szikla, G., Trottier, S., Vignal, J. P., Chauvel, P., Munari, C., & Chodkievicz, J. P. (1992). Surgical therapy for frontal epilepsies. *Advances in Neurology*, *57*, 707–732.

Tavano, A., Blohm, S., Knoop, C. A., Muralikrishnan, R., Scharinger, M., Wagner, V., Thiele, D., Ghitza, O., Ding, N., Menninghaus, W., & Poeppel, D. (2020). *Neural harmonics of syntactic structure* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.04.08.031575

Taylor, J. S. H., Rastle, K., & Davis, M. H. (2013). Can cognitive models explain brain activation during word and pseudoword reading? A meta-analysis of 36 neuroimaging studies. *Psychological Bulletin*, *139*(4), 766–791. https://doi.org/10.1037/a0030266

Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601.

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

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *ArXiv Preprint Physics/0004057*.

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. *2015 IEEE Information Theory Workshop (ITW)*, 1–5. https://doi.org/10.1109/ITW.2015.7133169

Toneva, M., Mitchell, T. M., & Wehbe, L. (2022). *Combining computational controls with natural text reveals new aspects of meaning composition* (p. 2020.09.28.316935). bioRxiv. https://doi.org/10.1101/2020.09.28.316935

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/749a8e6c231831ef7756db230b4359c8-Abstract.html

Tranel, D. (2009). The left temporal pole is important for retrieving words for unique concrete entities. *Aphasiology*, *23*(7–8), 867–884. https://doi.org/10.1080/02687030802586498

Tremblay, P., & Dick, A. S. (2016). Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain and Language*, *162*, 60–71. https://doi.org/10.1016/j.bandl.2016.08.004

Tripathy, J. K., Sethuraman, S. C., Cruz, M. V., Namburu, A., P., M., R., N. K., S, S. I., & Vijayakumar, V. (2021). Comprehensive analysis of embeddings and pre-training in NLP. *Computer Science Review*, *42*, 100433. https://doi.org/10.1016/j.cosrev.2021.100433

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

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, *311*(5761), 670–674.

Turing, A. (1950). *Computing Machinery and Intelligence*.

Tyukin, I., Gorban, A. N., Calvo, C., Makarova, J., & Makarov, V. A. (2019). High-Dimensional Brain: A Tool for Encoding and Rapid Learning of Memories by Single Neurons. *Bulletin of Mathematical Biology*, *81*(11), 4856–4888. https://doi.org/10.1007/s11538-018-0415-5

Urai, A. E., Doiron, B., Leifer, A. M., & Churchland, A. K. (2022). Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, *25*(1), Article 1. https://doi.org/10.1038/s41593-021-00980-9

Vandenberghe, R., Nobre, A. C., & Price, C. J. (2002). The Response of Left Temporal Cortex to Sentences. *Journal of Cognitive Neuroscience*, *14*(4), 550–560. https://doi.org/10.1162/08989290260045800

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., & Francart, T. (2018). Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. *Journal of the Association for Research in Otolaryngology*, *19*(2), 181–191. https://doi.org/10.1007/s10162-018-0654-z

Varoquaux, G., & Poldrack, R. A. (2019). Predictive models avoid excessive reductionism in cognitive neuroimaging. *Current Opinion in Neurobiology*, *55*, 1–6. https://doi.org/10.1016/j.conb.2018.11.002

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

Vyas, S., Golub, M. D., Sussillo, D., & Shenoy, K. V. (2020). Computation Through Neural Population Dynamics. *Annual Review of Neuroscience*, *43*, 249–275. https://doi.org/10.1146/annurev-neuro-092619-094115

Wada, J. (1949). A new method for the determination of the side of cerebral speech dominance: A preliminary report on the intracarotid injection of sodium amytal in man. *Igaku Seibutsugaku*, *14*, 221–222.

Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, *38*, 90–95. https://doi.org/10.1016/j.cobeha.2021.01.002

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), Article 6. https://doi.org/10.1038/s41593-018-0147-8

Warren, T., White, S. J., & Reichle, E. D. (2009). Investigating the causes of wrap-up effects: Evidence from eye movements and E–Z Reader. *Cognition*, *111*(1), 132–137. https://doi.org/10.1016/j.cognition.2008.12.011

Warstadt, A., & Bowman, S. R. (2022). *What Artificial Neural Networks Can Tell Us About Human Language Acquisition* (arXiv:2208.07998). arXiv. http://arxiv.org/abs/2208.07998

Wernicke, C. (1874). *Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis*. Cohn.

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

Whittington, J. C. R., & Bogacz, R. (2017). An Approximation of the Error Backpropagation Algorithm in a Predictive Coding Network with Local Hebbian Synaptic Plasticity. *Neural Computation*, *29*(5), 1229–1262. https://doi.org/10.1162/NECO_a_00949

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

Whittington, J. C. R., Warren, J., & Behrens, T. E. J. (2022, May 9). *Relating transformers to models and neural representations of the hippocampal formation*. International Conference on Learning Representations. https://openreview.net/forum?id=B8DVo9B1YE0

Willems, R. M., Nastase, S. A., & Milivojevic, B. (2020). Narratives for Neuroscience. *Trends in Neurosciences*, *43*(5), 271–273. https://doi.org/10.1016/j.tins.2020.03.003

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, *30*(5), 945–982.

Woolnough, O., Donos, C., Rollo, P. S., Forseth, K. J., Lakretz, Y., Crone, N. E., Fischer-Baum, S., Dehaene, S., & Tandon, N. (2020). Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nature Human Behaviour*, 1–10. https://doi.org/10.1038/s41562-020-00982-w

Wu, D. H., Waller, S., & Chatterjee, A. (2007). The functional neuroanatomy of thematic role and locative relational knowledge. *Journal of Cognitive Neuroscience*, *19*(9), 1542–1555. https://doi.org/10.1162/jocn.2007.19.9.1542

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

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), Article 3. https://doi.org/10.1038/nn.4244

Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525. https://doi.org/10.1016/j.tins.2011.06.006

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, *16*(8), Article 8. https://doi.org/10.1038/nrn3962

Zaccarella, E., & Friederici, A. D. (2015). Merge in the Human Brain: A Sub-Region Based Functional Investigation in the Left Pars Opercularis. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01818

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-11786-6

Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science (New York, N.Y.)*, *256*(5058), 846–849. https://doi.org/10.1126/science.1589767

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

Zhang, L., & Pylkkänen, L. (2018). Composing lexical versus functional adjectives: Evidence for uniformity in the left temporal lobe. *Psychonomic Bulletin & Review*, *25*(6), 2309–2322. https://doi.org/10.3758/s13423-018-1469-y

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

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3). https://doi.org/10.1073/pnas.2014196118