Neural Mechanisms of Linguistic Working Memory: Phrase Composition, Storage and Retrieval

Théo Desbordes (theo.desbordes@unige.ch)
Department of Basic Neurosciences, University of Geneva

Nicolas Piron (nicolas.piron@unige.ch)
Department of Basic Neurosciences, University of Geneva

Itsaso Olasagasti (miren.olasagasti@unige.ch)
Department of Basic Neurosciences, University of Geneva

Sophie Schwartz (sophie.schwartz@unige.ch)
Department of Basic Neurosciences, University of Geneva

Nina Kazanina (nina.kazanina@unige.ch)
Department of Basic Neurosciences, University of Geneva

Abstract

Understanding how the brain stores and manipulates linguistic information in working memory is central to understanding human cognition. Can we characterize the format of linguistic information storage in working memory? In this magnetoencephalography (MEG) study, participants read one-word, two-word, and five-word noun phrases followed by a matching task with a visual image. We found that individual word representations were maintained in neural activity for variable durations, depending on upcoming compositional demands. Critically, during a delay period following phrase reading, we observed a transition from word-specific to more abstract neural codes, with activity scaling alongside semantic complexity—suggesting compression of linguistic information. Retrieval dynamics revealed that access to surface-level properties was faster than to deeper semantic features, consistent with a decompression step. Finally, in ongoing work we explore potential contributions of reactivations —including coactivations and sequential replays— and oscillatory mechanisms such as phase-amplitude coupling, to the memory process. Together, these results map out the trajectory of linguistic processes, from online composition, through working memory storage, to retrieval. These findings place strong computational and biological constraints on models of linguistic working memory and could inform the design of new memory architectures in artificial conversational systems.

Keywords: Magnetoencephalography, Time-resolved decoding, Working memory, Language

Introduction

Humans uniquely possess the ability to bind successive words into novel, meaningful phrases. Yet, how the brain performs such composition—how individual word meanings are combined and represented in neural assemblies—remains an open question. Prominent computational theories, such as tensor-product representations (Smolensky, 1990), propose that phrases are encoded as vectorial structures that reflect both the meaning of individual words and the relations between them. These models exemplify factorized codes (Behrens et al., 2018), in which each component (e.g., word or syntactic role) can be recovered through linear operations. In contrast, compression is a general principle observed across cognitive domains, including auditory (Planton et al., 2021), and geometrical (Al Roumi et al., 2021) sequences. It suggests that the brain actively seeks compact representations that preserve meaning while minimizing redundancy. Under this hypothesis, the neural code for a phrase may no longer maintain linearly decodable traces of individual words. Instead, retrieval could require a decompression step, i.e., a specific operation applied to the memorandum that recovers the full representation. A core question, then, is whether linguistic phrases are stored as factorized representations—where individual word features remain linearly separable—or in a compressed form that integrates and reduces semantic redundancy. Additionally, working memory representations may be either active (sustained neural firing) (Goldman-Rakic, 1995; Leung et al., 2002) or silent (maintained via synaptic traces) (Stokes, 2015; Stokes et al., 2020). While silent mechanisms may suffice for passive storage, active neural patterns are likely required during composition and manipulation (Trübutschek et al., 2019), predicting distinct neural dynamics as phrases unfold. To tackle these questions, the present MEG study builds on previous work (Desbordes et al., 2024) that examined the neural instantiation of short noun phrases in working memory. In this dataset, participants read one-, two-, or five-word phrases describing colored shapes and judged whether a probe image matched the preceding phrase (Figure 1). Multivariate decoding was then applied to MEG signals to unravel the evolution of neural representations during three distinct phases: encoding, retention, and retrieval.

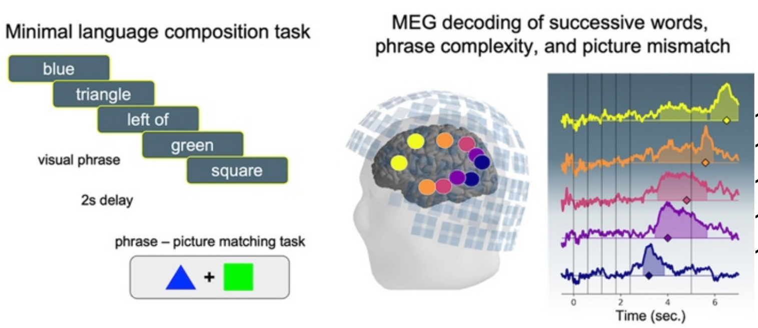

Figure 1: Study overview

Results

We trained logistic regression classifiers to decode individual words, separately for each category (e.g., one shape noun versus the other two), at each time point during the trial, yielding a time course of decoding performance. The decoding performance quickly rises after word presentation (Figure 2 Top) and is then maintained for longer when the word must be combined with upcoming words (not shown). The decoding performance then goes back to chance during the delay that precedes image presentation. To characterize neural activity during the delay period, we trained a linear regression model to predict a complexity score for each trial, based on the number of unique words in the phrase. Phrases containing entirely distinct words (e.g., “green circle right of red triangle”) were assigned a complexity score of 2, while those with maximal repetition (e.g., “blue square left of blue square”) received a score of 0. The regression model successfully predicts the phrases complexity all along the delay (Figure 2 Bottom). In additional analyses not included in this short manuscript, we show that neural activity during the delay period scales with this complexity measure, dissect the temporal dynamics of representations using temporal generalization, and demonstrate that retrieval is modulated by properties of the memoranda.

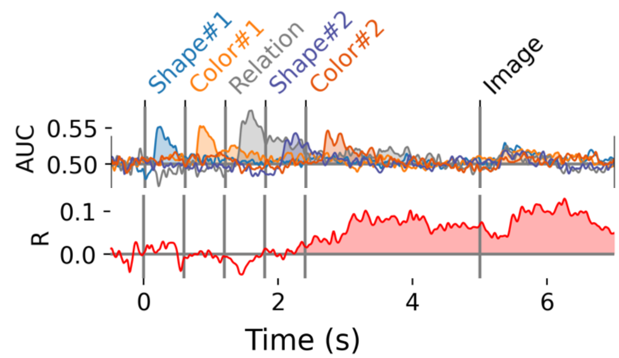

Figure 2: Decoding time courses
Top: decoding of individual words.
Bottom: regression decoding of the complexity of the sentence (number of unique words).

Discussion

Overall, our results support a compressed memory code. The storage format of phrases is such that individual properties are not linearly decodable, and computation is necessary to access all the information about the memorandum, akin to a decompression operation. While factorized codes have theoretical appeal (Bernardi et al., 2020) and are observed in nonhuman primates (Tian et al., 2024) and humans (Fan et al., 2025), they do not fit our MEG data. However, our results are compatible with other models of composition such as Vector-Symbolic

Architecture (Eliasmith & Anderson, 2003; Kleyko et al., 2022). Moving forward, we are currently extending this work along three major axes:

(1) the source localization of the identified effects, especially the compressed working memory code,

(2) the support of the memory trace by spontaneous reactivations, hypothesizing that the code is silent most of the time but reactivated intermittently, potentially with structure (e.g., sequential replay or coactivation of bound words), and

(3) testing whether the theta–gamma phase-amplitude coupling model of sequence memory (Heusser et al., 2016; Lisman & Idiart, 1995) applies to linguistic working memory: How many memory slots does a noun phrase occupy—one per word, or one in total due to compositional binding?

Acknowledgments

This research was funded by the National Center for Competence in Research Evolving Language (Swiss National Science Foundation Agreement #51NF40_180888, awarded to Nina Kazanina and Sophie Schwartz.

References

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

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

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

Desbordes, T., King, J.-R., & Dehaene, S. (2024). Tracking the neural codes for words and phrases during semantic composition, working-memory storage, and retrieval. *Cell Reports*, *43*(3). https://doi.org/10.1016/j.celrep.2024.113847

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

Fan, Y., Wang, M., Fang, F., Ding, N., & Luo, H. (2025). Two-dimensional neural geometry underpins hierarchical organization of sequence in human working memory. *Nature Human Behaviour*, *9*(2), 360–375. https://doi.org/10.1038/s41562-024-02047-8

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

Heusser, A. C., Poeppel, D., Ezzyat, Y., & Davachi, L. (2016). Episodic sequence memory is supported by a theta–gamma phase code. *Nature Neuroscience*, *19*(10), 1374–1380. https://doi.org/10.1038/nn.4374

Kleyko, D., Davies, M., Frady, E. P., Kanerva, P., Kent, S. J., Olshausen, B. A., Osipov, E., Rabaey, J. M., Rachkovskij, D. A., Rahimi, A., & Sommer, F. T. (2022). Vector Symbolic Architectures as a Computing Framework for Emerging Hardware. *Proceedings of the IEEE*, *110*(10), 1538–1571. Proceedings of the IEEE. https://doi.org/10.1109/JPROC.2022.3209104

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. Journal of Cognitive Neuroscience. https://doi.org/10.1162/08989290260045882

Lisman, J. E., & Idiart, M. A. (1995). Storage of 7 +/- 2 short-term memories in oscillatory subcycles. *Science (New York, N.Y.)*, *267*(5203), 1512–1515. https://doi.org/10.1126/science.7878473

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

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

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

Tian, Z., Chen, J., Zhang, C., Min, B., Xu, B., & Wang, L. (2024). Mental programming of spatial sequences in working memory in the macaque frontal cortex. *Science*, *385*(6716), eadp6091. https://doi.org/10.1126/science.adp6091

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