Semantic chunks save working memory resources: computational and behavioral evidence

Benjamin Kowialiewski (bkowialiewski@uliege.be)
Laboratoire de Psychologie et de NeuroCognition
Université Grenoble Alpes, Grenoble, France

Benoît Lemaire (benoit.lemaire@univ-grenoble-alpes.fr)
Laboratoire de Psychologie et de NeuroCognition
Université Grenoble Alpes, Grenoble, France

Sophie Portrat (sophie.portrat@univ-grenoble-alpes.fr)
Laboratoire de Psychologie et de NeuroCognition
Université Grenoble Alpes, Grenoble, France

Abstract

It is now well-established that long-term memory (LTM) knowledge, such as semantic knowledge, supports the temporary maintenance of verbal information in working memory (WM). This is for instance characterized by the recall advantage observed for semantically related (e.g. leaf - tree - branch) over unrelated (e.g. mouse - wall - sky) lists of items in immediate serial recall tasks. However, the exact mechanisms underlying this semantic contribution remain unknown. In this study, we demonstrate through a convergent approach involving computational and behavioral methods that semantic knowledge can be efficiently used to save attentional WM resources, thereby enhancing the maintenance of subsequent to-be-remembered items. These results have critical theoretical implications, and support models considering that WM relies on temporary activation within the LTM system.

Keywords: Working Memory; Computational Modeling; TBRS* model; Semantic Knowledge

Introduction

The influence of semantic knowledge on working memory (WM) performance is now supported by an increasing amount of studies. It has been shown that verbal items related at the semantic level (e.g. "leaf - tree - branch") are better recalled compared to verbal items sharing minimal characteristics at the semantic level (e.g. "mouse - wall - sky"), an effect also called *semantic relatedness* (Poirier & Saint-Aubin, 1995), suggesting a close interaction between long-term memory (LTM) knowledge and WM. Despite the extensive work done so far, the mechanisms responsible for these semantic influences remain poorly specified. In this study, we use a convergent approach involving computational modeling and behavioral experiments to test the hypothesis that semantic knowledge supports the temporary maintenance of verbal information, thereby saving WM resources that can be reallocated to maintain more information.

According to activation-based models of WM (Cowan, 1995), the maintenance of information over the short-term requires the temporary activation of LTM knowledge. As long as this information is kept sufficiently active, it can be accessed for subsequent recall. The influence of semantic knowledge on WM can be accounted for by considering that the locus of this activation in LTM lies within the linguistic system itself (Martin, Saffran, & Dell, 1996). As soon as a verbal item has to be maintained, it directly triggers the activation of associated phonological, lexical and semantic representations. The semantic relatedness effect can be explained by assuming that semantically related items, such as "leaf - tree - branch" reactivate each other, either via the semantic features they share (Dell, Schwartz, Martin, Saffran, & Gagnon, 1997), or via inter-item excitatory connections (Hofmann & Jacobs, 2014), which has the consequence to keep their activation level sufficiently high to be less susceptible to forgetting.

Although activation-based models make a good description of the interactions occurring in the LTM system, the way items are actually maintained in these models is strikingly lacking. *Attention-based models*, on the other side, make an excellent description of the way items are processed throughout the time course of WM. This is the case for instance as regards the Time-Based Resource Sharing Model (TBRS, Barrouillet, Bernardin, & Camos, 2004), which considers that WM performance is constrained by the balance between constantly decaying WM representations, and the time available to restore them through attentional refreshing. A computational implementation of this cognitive model, TBRS* (Oberauer & Lewandowsky, 2011), has shown to account for many important benchmark phenomena observed in WM tasks, such as cognitive load, serial position curves, omissions and transposition errors. The benefit of TBRS* is its ability to handle the dynamical aspects of human memory, by making a description of the functional mechanisms occurring at every stage (encoding, maintenance and recall) of WM processing. However, up to now, the TBRS* model has never been adapted in order to account for LTM influences on WM performance, which strongly limits its ability to make new predictions. Note that contrary to the TBRS* architecture which assumes decay as the cause of

forgetting in WM, other models such as SOB-CS (Oberauer, Lewandowsky, Farrell, Jarrold, & Greaves, 2012) consider instead that WM capacity is limited due to interference. This latter possibility is not considered in the present study for simplicity's sake, but we nevertheless do not deny the role of interference in WM.

Surprisingly, despite the evident complementarity of activation-based and attention-based models of WM, little attempt has been made in order to integrate them within a single formal architecture. The present study aims to integrate both approaches, by implementing some core principle assumed by activation-based models of WM within TBRS*, and test the ability of this new integration to account for data collected on human participants.

First, this new architecture should be able to reproduce the classical recall advantage observed for semantically related over unrelated items. By assuming that semantically related items constantly reactivate each other, their activation level should be much higher compared to semantically unrelated items which do not benefit from this strong co-activation. Due to this constant reactivation, semantically related items should be less susceptible to the deleterious effect of decay, leading to overall higher recall performance.

Second, and most importantly, semantic relatedness should save attentional WM resources that can potentially be reallocated to maintain more information. This *attentional resource saving hypothesis* stems from previous studies showing that when participants are invited to maintain in WM a set of letters, these letters are better recalled if preceded by a *chunk* (Thalmann, Souza, & Oberauer, 2018). For instance, in the target sequences "CLFVDHP" and "PDFLCHV", recall performance for "LCHV" is higher compared to "VDHP", because the former is preceded by an acronym (i.e. "PDF") which needs less refreshing episodes to be maintained, thereby leaving more free time available to counteract the deleterious effect of decay at the whole-list level. Similarly, the new architecture we propose predicts that if the target sequence "leaf - tree - branch - mouse - wall - sky" is to be maintained, there should be a recall advantage for "mouse - wall - sky" compared to a situation in which these three words were not preceded by semantically related items. The reason is that the semantic triplet members "leaf - tree - branch" would benefit from strong excitatory connections and would consequently require less refreshing episodes. Hence, there will be more free time available to refresh the semantically unrelated items "mouse - wall - sky", leading to better recall performance compared to the same items within lists composed of completely semantically unrelated items. This is what we mean by saving WM resources: semantically related items free up refreshing opportunities for the benefit of the remaining ones.

This second effect can be tested experimentally, and computational modeling offers the opportunity to investigate the underlying mechanisms in a fine-grained manner. It is a new prediction derived directly from the integration of both activation-based and attention-based models, that neither architecture considered alone is able to predict.

To sum up, we integrated within an attention-based WM model some principles derived from activation-based models to simulate the semantic relatedness effect. The output of this new model was compared to the recall performance of human participants performing the same experiment to test a new, still unobserved prediction derived from this model, i.e. to what extent semantic relatedness saves WM resources.

A new model integrating activated long-term memory and attention

The new architecture we present here is an adaptation of TBRS*, a functional computational WM model containing two layers: one coding for positional information and the other one coding for item information (see **Figure 1a**). In order to model LTM effects in TBRS*, the new architecture (see **Figure 1b**) includes a separate, LTM layer that is decoupled from the positional information in WM.

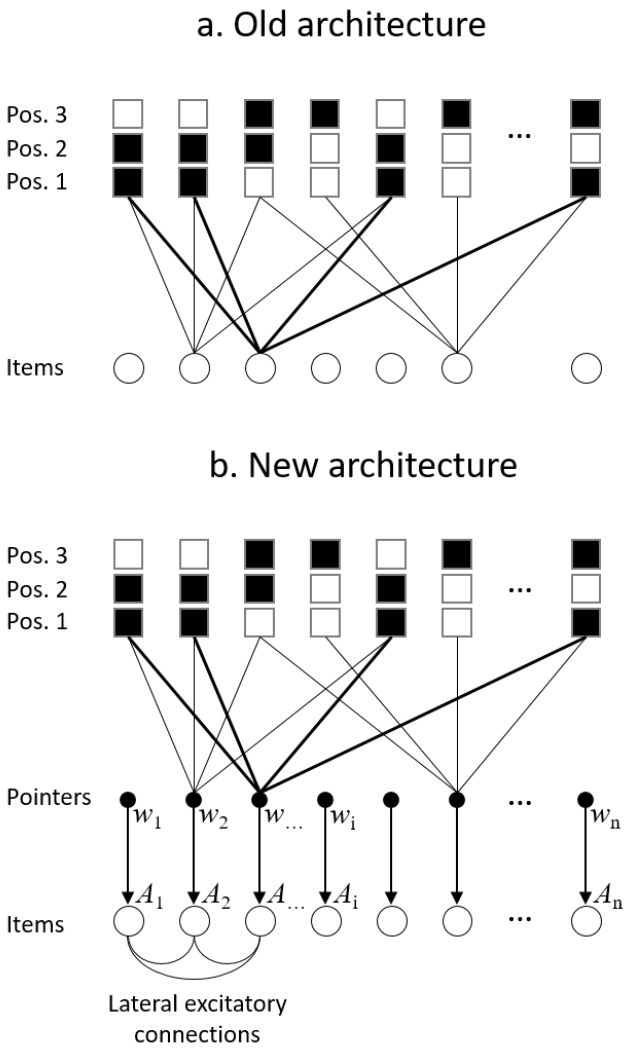

Figure 1. Illustration of the general (a) old and (b) new TBRS* architecture.

The use of a distinct LTM layer is motivated by three main reasons. The first one is mechanistic: without the possibility to store item information independently of positional information, the model would not be able to predict the recall advantage for semantically related over semantically unrelated items, because the co-activation of two related items will lead to their encoding within the same position. Due to this item-position co-occurrence effect, the model would predict a strong deleterious impact on the ability to recall serial order information. The second reason is theoretical: according to the embedded processes model (Cowan, 1995) which also frames the present study, the presentation of an item triggers its activation in the LTM knowledge base, and this activation is supposed to be independent from WM processes. The third and last reason stems from empirical evidence showing a dissociation between the ability to recall item and serial order information (Henson, Hartley, Burgess, Hitch, & Flude, 2003; Majerus, 2019).

The nodes that compose the LTM layer can be semantically related or not. For simplicity, semantic knowledge is not explicitly represented, but the mechanisms underlying the core principles behind the semantic relatedness effect are kept: items related at the semantic level are supposed to mutually activate each other within the LTM system. This mutual activation principle is modeled by including lateral excitatory connections between the nodes sharing a semantic relationship within the LTM layer. Once an item A_i receives a given amount of activation, all the semantically related items A_j also receive a proportion of this activation:

$$A_{j,t} = min(1, (A_{j,t-1} + A_{i,t-1}\lambda))$$

Where λ is the weight that connects A_i and A_j, and t refers to the timestamp of the ongoing iteration. The min function insures that the activation level will not exceed 1. This mechanism mimics the spreading activation phenomenon observed in semantic priming tasks (e.g. the presentation of "boat" preactivates "captain"). In addition, during the time the items decay, the spreading of activation continues to occur. In other words, they still receive an amount of activation after decay, such that:

$$\Delta A_i = (1 - A_i) \, . \, tanh(\lambda \sum A_{j,t-1})$$

Where the second factor is the normalization by the hyperbolic tangent of the total activation received by node A_i from all related nodes A_j. This way of representing semantic relationships by means of lateral excitatory connections in an all-or-none fashion was sufficient to describe the WM mechanisms involved. This could be extended to take into account various degrees of semantic relationships instead of only one.

The position of each to-be-remembered item is represented in a distributed fashion using positional markers. Adjacent positions are assumed to share some degree of overlap. Serial order information is kept in memory by associating the positional markers to a set of nodes acting as pointers towards items. The role of the pointers is to index the LTM representations to which they point to. Modeled this way, memory for serial order information will not be affected by the spreading of activation that occurs within the LTM layer. As previously mentioned, dissociating the activation in LTM from serial order processes is an important theoretical choice, because empirical evidence show that LTM knowledge minimally impacts the ability to maintain serial order information (Majerus, 2019). This new architecture contrasts with the old TBRS* architecture which considered that the only information that is stored is the item-to-position association, but the core WM functioning remains almost unchanged.

Encoding. Encoding is performed by activating the current item in the LTM layer. The other semantically related items receive a portion of this activation via their connections. At the same time, the pointers-to-positions associations are also created following a simple Hebbian learning rule.

Maintenance. Following the encoding phase, the model enters in a dynamic balance state constrained by two opposed phenomena: decay and refreshing. The WM representations are constantly decaying, unless they can be refreshed using the focus of attention, a central bottleneck limited to one item. To keep the model simple, we assumed that decay only affects item information. When attention is available, items are constantly refreshed by a rapid switching of the focus of attention from one item to another. During refreshing, WM representations are reinforced using the same principles as those used during the encoding stage. Note that spreading of activation in LTM also occurs during refreshing.

The rate of refreshing has been set to 80 ms per item as in the original TBRS* model. There are controversies as regards whether the focus of attention refreshes the items in a cumulative way or using a different schedule (Vergauwe et al., 2016). In this study, we assumed that the human cognitive system is efficient, and that participants try to optimize their available resources as much as possible. Hence, refreshing operates in priority over the item that is the most likely to be forgotten, a mechanism called *Least Activated First* (Lemaire, Pageot, Plancher, & Portrat, 2018).

Retrieval & recall. Before being refreshed and/or recalled, an item must be first retrieved. In the model, retrieval is performed by feeding the network via the positional markers for a given position, and then selecting the item the most associated to that position. This selection is constrained by the product of two sets of information: the evidence w_i accumulated within the pointers after feeding the positional markers, and the strength of activation a_i in LTM. For each atomic 80 ms refreshing step, the least activated item in LTM is selected and refreshed.

Recall is performed by retrieving the items via the positional markers one by one. After each recall episode, the WM representations continue to decay. In addition, a response suppression mechanism is implemented to avoid constant repetition of the same item (Lewandowsky, 1999). The response suppression mechanism consists in pushing the weights that connect the positional markers to the pointers towards negative values.

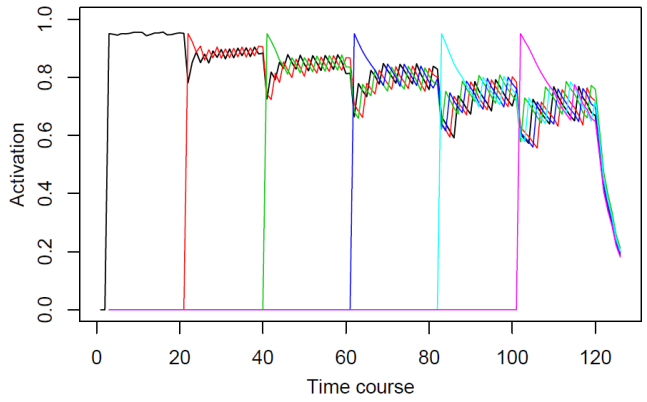

Figure 2. Time course of the item activation over one trial.

In the model, errors are caused by three main sources: (a) the overlap between positional markers, (b) a noise parameter, and (c) a retrieval threshold. Transposition errors (e.g., recalling "ACB" instead of "ABC") are caused by (a) and (b), and item errors such as extra-list intrusions (i.e., an item not presented in the list) and omissions (i.e., complete forgetting) are mainly caused by (b) and (c). An example of time course of a 6-item list is displayed in **Figure 2**.

Experiment

In this study, we tested the *attentional resource saving* hypothesis on a memory task involving three different semantic conditions, in which 6 words had to be maintained and immediately recalled:

- A first condition in which the first half of the to-be-remembered list was composed of items from the same semantic category (**C1**; Semantic Chunk in first half, e.g., leaf - tree - branch - mouse - wall - sky).
- A second condition where only the second half of the to-be-remembered list was composed of items from the same semantic category (**C2**; Semantic Chunk in second half, e.g. cloud - wolf - mud - hand - arm - leg).
- A third condition in which all the items were drawn from a different semantic category (**NC**; Non-Chunked).

Overall, we predicted a recall advantage for semantically related over unrelated items, as classically observed. Critically, following the attentional resource saving hypothesis, there should be a recall advantage over positions 4, 5 and 6 in the C1 condition, and this compared to the same positions in the NC condition. The C2 condition was added in order to explore whether the position of the chunk is of critical importance. Previous studies have indeed shown that chunking saves WM resources, but only when the chunk is located at the beginning of the list (Thalmann et al., 2018).

Human participants

Material. We chose 120 words (one to three syllables long), drawn from forty different semantic categories, with three words per category. All the stimuli were recorded by a French native male speaker in a neutral voice.

The semantic categories were used in order to create the three different semantic conditions (N_{trial} = 20 in each condition C1, C2 and NC). The semantically unrelated lists were created by directly combining the words from the semantic categories, such that each word within a list could not share an obvious semantic relationship with another word in the list.

Procedure. Each of the six items were auditorily presented at a pace of 2 seconds per item. After the last item had been presented, the participants were invited to recall out loud the sequence in the order in which the items had been presented, and to substitute any item they could not remember with the word "blank". For both the simulations and the data collected on human participants, we used a strict serial recall criterion, whereby an item was considered correct only if recalled at the correct position. All the participants received the three semantic conditions in a fully within-subject design.

Statistical analysis. We performed a Bayesian analysis, in which evidence for H_1 and H_0 can be simultaneously tested, by directly comparing the alternative hypothesis against the null hypothesis, and vice versa. Evidence in favor of H1 is indicated by BF_{10}, and evidence in favor of H_0 is indicated by $1/BF_{10} = BF_{01}$. A BF of 1 provides no evidence, $1 < BF < 3$ provides anecdotal evidence, $3 < BF < 10$ provides moderate evidence, $10 < BF < 30$ provides strong evidence, $30 < BF < 100$ provides very strong evidence and $100 < BF$ provides extreme/decisive evidence. These labels serves only an indicative purpose, as the BF relies on continuous values of evidence rather than on arbitrary thresholds.

Results. Data were collected from thirty participants, aged between 18 and 33 (M = 20, SD = 3) with no history of neurological disorder or learning difficulties.

We first assessed recall performance as a function of the semantic condition (C1, C2, NC) and serial position (1 through 6). Using a Bayesian Repeated-Measures ANOVA, we found decisive evidence supporting the two main effects of semantic condition ($BF_{10} > 100$) and serial position ($BF_{10} > 100$). As can be seen in **Figure 3a**, there was a gradual recall performance increase as a function of semantic condition: C1 (M = .775) > C2 (M = .708) > NC (M = .645).

Importantly, the interaction between the semantic condition and serial position was also associated with decisive evidence ($BF_{10} > 100$). As can be clearly seen in **Figure 3a**, the impact of the different semantic conditions was not equivalent across all positions. This interaction was further explored using Bayesian paired-samples T-Tests.

When the semantic chunk was presented in positions 1, 2 and 3, there was a clear recall advantage over these positions compared to the non-chunked condition (i.e. C1 vs. NC; $BF_{10} > 100$, d = 1.199). Similarly, when the semantic chunk was presented in positions 4, 5 and 6, there was also a recall advantage over these positions compared to the unrelated condition (i.e. C2 vs. NC; $BF_{10} > 100$, d = 1.678). These results indicate that the classical semantic relatedness effect was replicated; there was better recall performance for semantically related items compared to semantically unrelated items.

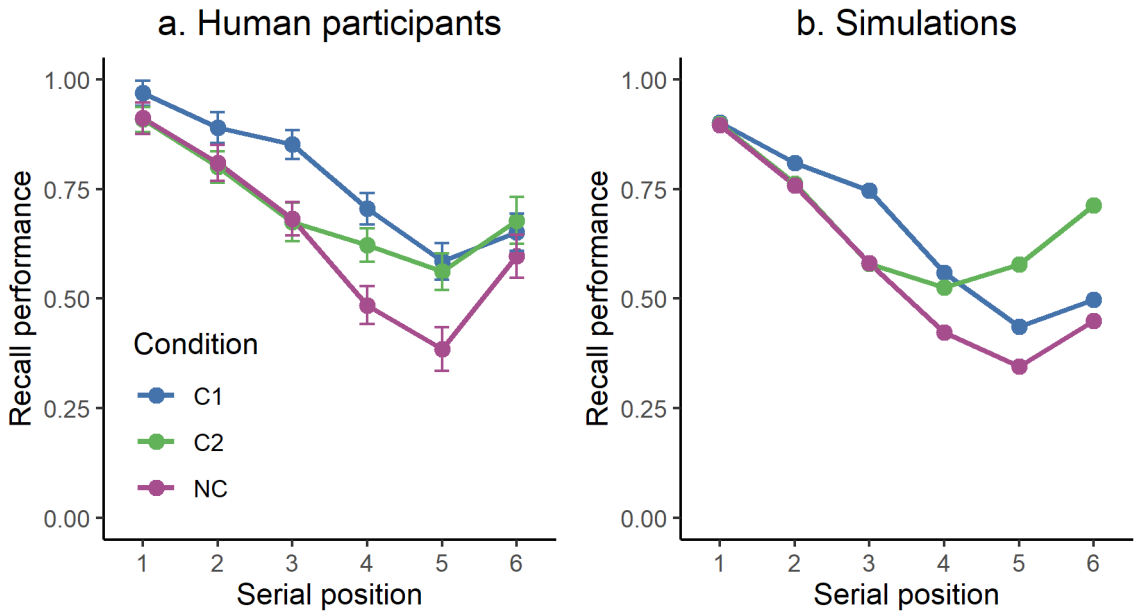

Figure 3. Recall performance across serial positions as a function of the semantic relatedness conditions (C1, C2, NC) for (a) human participants and (b) the computational model. Error bars indicate 95% confidence intervals, after correction for between-subject variability.

Critically, the attentional resource saving hypothesis predicts a recall advantage over positions 4, 5 and 6 when the semantic chunk is presented at the beginning of the list. This is indeed what we observed (C1 vs. NC; BF_{10} > 100, d = 1.429), indicating that the semantic chunk did indeed save WM resources.

Whether there could be a beneficial effect over positions 1, 2 and 3 when the semantic chunk is presented in the second half of the list is quite difficult to predict following the attentional resource saving hypothesis. The results indicate an absence of recall advantage, which was supported by moderate evidence (C2 vs. NC; BF_{10} = .307; BF_{01} = 3.259, d = -.183), showing that semantic chunks saved WM resources, but mainly when presented at the beginning of a to-be-remembered list.

Hence, the predictions derived from the attentional resource saving hypothesis were met: the semantic chunk had a beneficial effect on recall performance for subsequent, non-chunked items. However, the exact mechanisms responsible for the pattern of results we observed have yet to be specified in a formal manner. In the next section, we therefore used computational simulations in order to assess whether the mechanisms we supposed to be responsible for the attentional resource saving hypothesis are making the correct predictions.

Simulations

Simulation details. The model and the human participants performed exactly the same experiment, with 6 items presented at a pace of 2,000 ms per item. We assumed an encoding time of 500 ms, with an inter-item interval of 1,500 ms. All three experimental conditions (C1, C2 and NC) were tested on the model, the semantic relatedness being implemented via excitatory connections as described above.

Parameter estimation. Parameters of the model were estimated using a grid search method that explored 13,376 different parameter combinations. The list of parameters used and associated range of values are displayed in the upper part of **Table 1**. Each set of parameters was estimated using 1,000 simulations in the neutral condition (i.e. NC) only. The fit of the model was then assessed via direct comparison with the empirical data using the Root Mean Squared Error (RMSE), and the configuration of parameters that minimized the error was selected.

After the best set of parameters had been selected, we then performed a new search over the strength of semantic connections λ in order to find a value that minimized the error against the behavioral data in terms of mean difference between conditions C1 and NC over positions 1, 2 and 3.

Results. The best set of parameters that minimized the error over the NC condition were associated with a RMSE value of .075 (see **Table 1** for the associated parameters). The results of this model are displayed in **Figure 3b**.

The model was able to capture the classical recall advantage for semantically related over unrelated words (i.e. C1 > NC in positions 1 through 3, and C2 > NC in positions 4 through 6). Most importantly, the model also produced a recall advantage over positions 4 through 6 in C1 compared to NC, indicating that the semantic chunk at the beginning of the list did indeed save WM resources. This recall advantage was naturally found without directly fitting the model based on this pattern of result. Interestingly, no obvious similar

Table 1. Range of values explored within the grid search. Note that the lambda has been estimated separately.

Parameter	Meaning	Minval	Maxval	Steps	Best
p	Overlap between positions	0	.9	.05	.55
σ	Noise added at retrieval	0	.1	.01	0
θ	Retrieval threshold	0.05	.4	.05	.25
D	Decay rate	.1	.8	.1	.4
λ	Lateral connections value in LTM	.01	.05	.001	.017

recall advantage over positions 1 through 3 was observed in C2 compared to NC, suggesting that the within-list position of the chunk does matter in order to save WM resources.

The attentional resource saving hypothesis predicted that the recall advantage on positions 4, 5, 6 when items 1, 2, 3 are semantically related is due to more refreshing opportunities in C1 compared to NC. Indeed, because the items of a semantic chunk are overall associated with greater activation values and decay more slowly than unrelated items, they need less refreshing episodes, thus leaving more attentional resources available. To test this prediction in the most direct manner, we computed the average number of refreshing episodes for each item. Results over 10,000 simulations indicate that items 1, 2 and 3 were refreshed 15,861 times less in the C1 condition compared to the NC condition (out of about 1,111,000 refreshing episodes), indicating a smaller number of refreshing episodes over the semantic chunk. This smaller number of refreshing episodes over the semantic chunk actually shifted toward items 4, 5 and 6: these items were refreshed 16,190 times more in the C1 condition compared to the NC condition. Critically, the same pattern was observed in C2, but was smaller compared to C1: the shift of refreshing episodes was only about 10,000. Hence, the introduction of a semantic chunk did indeed change the pattern of refreshing episodes.

To sum up the results of this study, there was a recall advantage for items that directly followed a semantic chunk, and this compared to a condition where no semantic chunk was present. In contrast, no such advantage for items preceding a chunk was observed. This pattern of results was found both in the behavioral and computational simulations. In the next section, we discuss the theoretical implications of these results.

Discussion

In this study, we investigated the fundamental processes underlying the semantic relatedness effect, by testing the hypothesis that semantic knowledge can be used to save WM resources. We observed that when a semantic chunk composed of semantically related items was presented at the beginning of a to-be-remembered list, WM resources were saved, as indicated by a recall advantage over the remaining items of the list. Moreover, the computational modeling approach allowed us to formally establish the plausibility that this gain was characterized by more refreshing episodes over the end of the list when the beginning of the list was semantically related.

The computational architecture we used is a hybrid connectionist model integrating both the maintenance processes operating during WM processing (Barrouillet et al., 2004), and the activation in the LTM system (Cowan, 1995; Martin et al., 1996). More specifically, as in TBRS* (Oberauer & Lewandowsky, 2011), the constantly decaying WM representations need to be maintained via attentional refreshing, and the semantically related items in LTM mutually reactivate each other (Dell et al., 1997; McClelland & Rumelhart, 1981).

The overall recall benefit for semantically related over semantically unrelated items suggests that the mechanisms we implemented to model semantic effects are plausible. At the same time, it appears that when the semantic chunk was presented at the end of the list, the model produced a semantic relatedness effect that was quite unrealistic compared to what is actually observed in the empirical data. This latter aspect shows that there is still room for improvement to capture semantic effects in a more general manner.

The innovative outcome of this model is its ability to predict the recall advantage for the subsequent, semantically unrelated items when a semantic chunk was present in the list. Because semantically related items benefit from strong co-activations, they need less refreshing episodes, leaving more free time available to refresh the other, non-related items of the list. In fact, this behavior was well captured by the refreshing schedule produced by the model, with more refreshing episodes observed for items that directly followed a semantic chunk. This behavior is partially explained by the Least Activated First principle: items that are the most likely to be forgotten have a priority status during the refreshing process (Lemaire et al., 2018). Hence, since the system is constantly trying to optimize the resource allocation, there is no reason to refresh “leaf - tree - branch” because, thanks to their semantic relatedness, their activation level is higher compared to the other, semantically unrelated items.

Interestingly, the model predicted an absence of recall advantage for items in positions 1 through 3 when the semantic chunk appeared in positions 4 through 6, and this absence of recall advantage was similarly observed among human participants. This is due to the fact that the opportunity to save attentional resources happens much later when the semantic chunk is presented in the second half of the list, which does not provide enough boost to increase recall of the first items of the list. In contrast, when the semantic chunk is

presented in the first part of the list, semantically related items are already fully co-activated, and does not require much refreshing episodes throughout the rest of the trial.

An obvious alternative explanation to account for the empirical results observed could be that participants only maintained the supra-ordinate semantic category from which the related items belong (Martin, Minkina, Kohen, & Kalinyak-Fliszar, 2018). For instance, after the presentation of "leaf - tree - branch", the participants might simply maintain the conceptual unit "forest", and use it as a cue at the moment of retrieval. Since we did not compare both accounts in a formal computational implementation, it is difficult to rule out this possibility. However, such a mechanism has already been successfully implemented to simulate experiments in which participants had to maintain chunks composed of letters. A semantic chunking mechanism is therefore likely to lead to the same overall conclusions (Portrat, Guida, Phénix, & Lemaire, 2016).

To sum up, this study demonstrates the potentiality of a new architecture integrating the supports of attentional and LTM in WM functioning. By considering that semantically related items reactivate each other within the LTM system, we have shown that attentional WM resource can be saved thanks to this constant reactivation. Importantly, this study demonstrates the whole complexity of the interactions occurring between LTM and attention when maintenance over the short-term is required.

References

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, *133*(1), 83–100. https://doi.org/10.1037/0096-3445.133.1.83

Cowan, N. (1995). *Attention and memory: An integrated framework*. Oxford, England: Oxford University Press.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical Access in Aphasic and Nonaphasic Speakers. *Psychological Review*, *104*(4), 801–838.

Henson, R. N. A., Hartley, T., Burgess, N., Hitch, G., & Flude, B. (2003). Selective interference with verbal short-term memory for serial order information : A new paradigm and tests of a timing-signal hypothesis. *The Quarterly Journal of Experimental Psychology*, *56A*(8), 1307–1334. https://doi.org/10.1080/02724980244000747

Hofmann, M. J., & Jacobs, A. M. (2014). Interactive activation and competition models and semantic context: From behavioral to brain data. *Neuroscience and Biobehavioral Reviews*, *46*(P1), 85–104. https://doi.org/10.1016/j.neubiorev.2014.06.011

Lemaire, B., Pageot, A., Plancher, G., & Portrat, S. (2018). What is the time course of working memory attentional refreshing ? *Psychonomic Bulletin & Review*, *25*(1), 370–385. https://doi.org/10.3758/s13423-017-1282-z

Lewandowsky, S. (1999). Redintegration and Response Suppression in Serial Recall: A Dynamic Network Model. *International Journal of Psychology*, *34*(5/6), 434–446.

Majerus, S. (2019). Verbal working memory and the phonological buffer: The question of serial order. *Cortex*, *112*(May), 122–133. https://doi.org/10.1016/j.cortex.2018.04.016

Martin, N., Minkina, I., Kohen, F. P., & Kalinyak-Fliszar, M. (2018). Assessment of linguistic and verbal short-term memory components of language abilities in aphasia. *Journal of Neurolinguistics*, *48*(December 2017), 199–225. https://doi.org/10.1016/j.jneuroling.2018.02.006

Martin, N., Saffran, E. M., & Dell, G. S. (1996). Recovery in deep dysphasia: Evidence for a relation between auditory–verbal STM capacity and lexical errors in repetition. *Brain and Language*, *52*(1), 83–113.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*(5), 375.

Oberauer, K., & Lewandowsky, S. (2011). Modeling working memory: a computational implementation of the Time-Based Resource-Sharing theory. *Psychonomic Bulletin & Review*, *18*(1), 10–45. https://doi.org/10.3758/s13423-010-0020-6

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review*, *19*(5), 779–819. https://doi.org/10.3758/s13423-012-0272-4

Poirier, M., & Saint-Aubin, J. (1995). Memory for related and unrelated words: Further evidence on the influence of semantic factors in immediate serial recall. *The Quarterly Journal of Experimental Psychology*, *48*(2), 384–404. https://doi.org/10.1080/14640749508401396

Portrat, S., Guida, A., Phénix, T., & Lemaire, B. (2016). Promoting the experimental dialogue between working memory and chunking: Behavioral data and simulation. *Memory and Cognition*, *44*(3), 420–434. https://doi.org/10.3758/s13421-015-0572-9

Thalmann, M., Souza, A. S., & Oberauer, K. (2018). How Does Chunking Help Working Memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Vergauwe, E., Hardman, K. O., Rouder, J. N., Roemer, E., Mcallaster, S., & Cowan, N. (2016). Searching for serial refreshing in working memory: Using response times to track the content of the focus of attention over time. *Psychonomic Bulletin & Review*, *23*(6), 1818–1824. https://doi.org/10.3758/s13423-016-1038-1