UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Thinking through syntax: Expanding the scope of "thinking for speaking"

Permalink
https://escholarship.org/uc/item/0w4161w4

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Xue, Yuyan
Culbertson, Jennifer

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Thinking through syntax: Expanding the scope of "thinking for speaking"

Yuyan Xue (yx324@cam.ac.uk)
Department of Theoretical and Applied Linguistics, University of Cambridge
Department of Linguistics, University of Hong Kong

Jennifer Culbertson (jennifer.culbertson@ed.ac.uk)
School of Philosophy, Psychology and Language Sciences, University of Edinburgh

Abstract

The "thinking for speaking" hypothesis proposes that our language can influence cognition during language production or interpretation, directing our attention to the grammatical and/or semantic categories readily codable in our language. Beyond the codability of grammatical and semantic categories, the role of syntactic hierarchy–a core feature of human language—has not been studied so far. This study addresses this gap by investigating the effect of learning different complex noun phrase (NP) structures on English native participants' similarity judgments of objects which differed in color or number. In Experiment 1, as the training proceeded, participants who learned to describe novel objects following an unusual NP structure that highlights the dimension of number over color were more likely to judge objects matching on number; by contrast the judgments of participants who learned a more typical NP structure that highlights color over number did not change significantly over time. The training-specific effect observed in Experiment 1 failed to emerge in Experiment 2 where online language involvement was reduced. These results extend the scope of "thinking for speaking", suggesting that hierarchical structure in syntax may also influence cognition during language use. They also shed light on the potential for cognitive flexibility in representations of the NP.

Keywords: thinking for speaking; syntactic hierarchy; noun phrase; similarity judgment

Introduction

Thinking for speaking

The interplay between language and cognition has been the subject of debate at least since Aristotle (see Aristotle's *On Interpretation*). Among various proposals, the universalist account proposes that our cognition is unaffected by our language (e.g., Pinker, 1994). Whilst the "thinking for speaking" hypothesis (Slobin, 1996, Slobin, 2004) postulates that during language production or processing, our language may channel our attention to elements of experience readily codable in our language. As languages vary in the codability of different grammatical and/or semantic categories, speakers of different languages may vary to some degree in how they perceive or represent the world when producing or interpreting language.

Empirically, the "thinking for speaking" hypothesis has been studied in a number of domains including grammatical gender (Samuel, Cole, & Eacott, 2019), space (Choi, McDonough, Bowerman, & Mandler, 1999), evidentiality (Ünal & Papafragou, 2020), and event cognition (Vanek, 2020). Event cognition is the most well-studied domain, where the focus has been on cross-linguistic differences in the encoding of manner and path information. Based on Talmy (2000)'s typology, satellite-framed languages (e.g., English) prefer to encode the manner in the main verb (e.g., S/he *ran* up to the third floor) and the path in a verb satellite (e.g., S/he ran *up to the third floor*); whereas verb-framed languages (e.g., Japanese) tend to encode path in the main verb and often drop manner information (e.g., 彼は三階に上がった ; translation: S/he *ascended* to the third floor; manner information omitted). This is because verb-framed languages often do not have an obligatory syntactic slot to encode manner. In other words, languages of the two types differ in the codability of the semantic category of manner. In line with this typological difference, after describing a set of stimuli, English monolinguals categorized motion events on the basis of manner significantly more than Japanese monolinguals (Wang & Li, 2023; see also Soroli, 2024 for more mixed results; Bylund & Athanasopoulos, 2024 for review on "thinking for speaking" in motion events).

In contrast to other hypotheses, such as the salience hypothesis (see Papafragou & Selimis, 2010) or the linguistic relativity hypothesis (Whorf, 1956), the "thinking for speaking" hypothesis predicts that language effects on cognition are restricted to situations where (the preparation of) language production or interpretation is involved. Indeed, numerous studies have found language effects on cognition disappear or are substantially reduced when online recruitment of language is no longer encouraged or is disrupted through verbal interference (e.g., repeating digits or syllables while performing the task measuring cognition) (Papafragou & Selimis, 2010; Vanek & Selinker, 2017; Hickmann, Engemann, Soroli, Hendriks, & Vincent, 2017).

Syntactic hierarchies and cognition

Existing studies on "thinking for speaking" have focused on words, morphemes, and/or grammatical properties such as gender and aspect. More complex syntactic representations have not been studied from this perspective. Representations of syntax, for example structural hierarchies, have been argued to be a core and potentially unique aspect of human language (e.g., Friederici, 2017, although see Ferrigno, Cheyette, Piantadosi, & Cantlon, 2020). However, not all approaches to syntax agree on the role of abstract representations in language learning and use (e.g., Tomasello, 2003, Goldberg, 2003, Frank, Bod, & Christiansen, 2012). It is

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

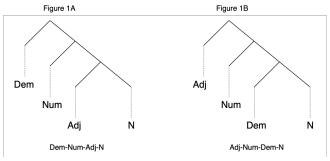

Figure 1: Two possible NP hierarchies. Following Martin et al. (2020), Figure 1A is common and homomorphic, Figure 1B is uncommon and non-homomorphic

therefore worth investigating whether "thinking for speaking" effects can also be driven by differences at this level of representation.

Importantly, to test the effect of syntactic structure, we must deconfound it from other possible effects, like the codability of semantic categories. Here, as a test case, we use complex noun phrase (NP) word orders, and the hierarchical structure(s) such word orders implicate. A noun (N) can be modified by adjectives (Adj, e.g., white), numerals (Num, e.g., two), and demonstratives (Dem, e.g., this, that) at the same time. Human languages vary in the order in which they arrange these three types of modifiers, but some orders are more common than others. This has led to the proposal that noun phrase structure across languages tends to follow a common syntactic hierarchy (Cinque, 2005; Abels & Neeleman, 2012), as in Figure 1A (from Martin, Holtz, Abels, Adger, & Culbertson, 2020). In this hierarchy, adjectives combine with the noun to create a semantically unified sub-constituent, numerals specify the quantity of this constituent to make it a countable unit, and demonstratives connect this unit to the pragmatic context, such as indicating its location relative to the speaker (Culbertson & Adger, 2014). The two most common orders in the world's languages, N-Adj-Num-Dem as in Thai, and Dem-Num-Adj-N as in English ("these two white cats") perfectly reflect this hierarchy. An additional six languages can be transparently derived from it (e.g., Dem Num N Adj, by linearising N before only Adj). In contrast, only a small proportion of languages exhibit an order that does not transparently reflect this hierarchy. For example, some Bantu languages, like Kîîtharaka have N-Dem-Num-Adj (Martin, Adger, Abels, Kanampiu, & Culbertson, 2024). Virtually no language exhibits the order Adj-Num-Dem-N (Greenberg, 1963; Cinque, 2005; Dryer, 2018). These languages either involve more radical departures from the hierarchy, or implicate a different hierarchy altogether as in Figure 1B.

Evidence for a preference for orders that conform to the hierarchy in Figure 1A—which we call 'homomorphic' following Martin, Abels, Ratitamkul, and Culbertson (2019)—comes from miniature artificial language learning studies. In these studies, speakers of English (Dem-Num-Adj-N), Thai (N-Adj-Num-Dem), and Kîîtharaka (N-Dem-Num-Adj) were taught a set of artificial lexical items for nouns, and modifiers, and were also taught a non-native-like position for the nouns (e.g., post-nominal in English; pre-nominal in Thai and Kîîtharaka). They were given no evidence for the relative order of modifiers in the artificial language. While English and Thai-speakers have a native language order that is homomorphic to the hierarchy in Figure 1A, Kîîtharaka speakers do not. Nevertheless, participants in all three populations preferred to arrange the artificial words in a homomorphic order—with the adjective closest to the noun, and the demonstrative farthest away (Culbertson & Adger, 2014; Culbertson, Schouwstra, & Kirby, 2020; Martin et al., 2024). The authors took this finding, together with the theories above, as suggesting a universal hierarchical representation, and a preference for orders conforming to it, regardless of native language experience. In other words, they argue that, in this case, there is no apparent effect of language on how humans hierarchically represent complex NPs.

Interestingly though, Culbertson et al. (2020) argue that the hierarchy in Figure 1A may, at some level, reflect differences in conceptual closeness between objects and the kinds of properties encoded by adjectives, numerals, and demonstratives. They use cross-linguistic corpus evidence to infer that on average there is higher mutual information between objects and adjectival properties (e.g., color) than between objects and their numerosities, with the lowest mutual information between objects and demonstrative-like properties (e.g., location relative to the speaker). They argue that humans may prefer a transparent way of mapping conceptual representations to a syntactic hierarchy, resulting in the observed ordering preferences. This suggests that, all things equal, if transparency between levels of representation is only a preference, experience with alternative orders *could* in principle lead to a restructuring of syntactic representations. Further, it suggests that there could be a relationship between syntactic structure and how people represent objects in the world. However, the artificial language learning tasks described above do not directly assess whether speakers perceive or represent complex objects differently depending on the syntactic structure of their native language.

One potential way to probe this more directly would be to design a task similar to the experiments on "thinking for speaking" described above. In such a task, we would compare two groups of speakers, e.g., English and Kîîtharaka, to see whether the descriptions they provide in their native language predicted differences in a perception or memory task with non-verbal stimuli. Here, we chose instead to use a randomized-controlled artificial language learning design. We did so because this kind of design can maximally control for confounds suffered by quasi-experimental, cross-linguistic comparative studies. Most obviously, in the latter type of experiment, cognitive differences can result from cultural rather than linguistic diversity (Casasanto, 2008). By contrast, researchers have argued that brief language training is analogous to a highly condensed version of language learning in the real world, without these potential confounds (Casasanto, 2008; Montero-Melis, Jaeger, & Bylund, 2016).

Experiment 1

To test the potential link between language and conceptual representations, here we use an artificial language learning experiment combined with a similarity judgment task of the kind featured in previous studies on "thinking for speaking".In Experiment 1, we taught participants either a homomorphic order, which conforms to the preferred hierarchy (e.g., Adj closest to N), or a non-homomorphic order, which does not conform to this hierarchy (e.g., Adj farthest from N). We ask whether describing objects using the non-conforming order can impact participants' perceptions of similarity of those objects. Specifically, we trained English native speakers to describe unfamiliar objects with artificial words using N-Adj-Num order (homomorphic control group) or a N-Num-Adj order (non-homomorphic experimental group).[1] Here, the adjectives all described colors of objects. After describing each image (the target) in the correct order for their group, participants were asked to judge which one out of two alternative images was more similar to the target. The two alternative images differed from the target image in either the color or the number of the set of pictured objects.

If "thinking for speaking" extends to the hierarchical information present in complex NPs, then we predict that learning to describe objects with different NP structures will lead to differences in participants' similarity judgments. In particular, because N and Num are closer in the non-homomorphic order, this implies a tighter structural relationship between N and Num. This should lead to an increased perception of similarity between images that share a numerosity.

Participants

Eighty-nine adult English native speakers were recruited via Prolific and randomly assigned to the experimental (N = 43, Mage = 35.95, 25 females) or the control group (N = 46, Mage = 39.22, 21 females). No participant reported knowing any non-homomorphic language.

Stimuli

Novel objects: To reduce the influence from participants' background knowledge, we chose three unfamiliar objects (#2005, #2013, #2022) from the NOUN database (Horst & Hout, 2016) which had very low average ratings of familiarity and name-ability (according to Horst & Hout, 2016). We used Adobe Photoshop to change their original colors into red and yellow. We made the color difference subtle (but easily distinguishable) to avoid ceiling effects and allow for more room for the potential group difference. The objects were all presented to the participants on a table with a cartoon girl sitting at the table (following Martin et al. (2020)).

[1] While the study with Kîîtharaka-speakers reported in Martin et al. (2024) used adjectives and demonstratives, previous studies with both English and Thai speakers also tested learners on numerals and adjectives (e.g., Martin et al., 2020). We chose to use this combination of modifiers, since they are easier to visually represent. Distance from the speaker requires more complex stimuli, and may be partially conflated with other properties such as whether the object appears on the left-hand or right-hand side of the screen, etc.

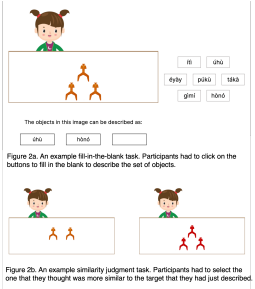

Figure 2: An example trial of Phase 5 for the experimental group

Novel words: The three novel objects (#2005, #2013, #2022) were *éyày*, *úhù*, and *ítì* respectively; red and yellow were *púkù* and *tákà*; numbers two and three were *gìmí* and *hònó*. All novel words were taken from Martin et al. (2020) except that we changed *hìmí* into *gìmí* to make the words for two and three more distinctive (more closely parallel to *púkù* and *tákà*).

Procedure

Following Martin et al. (2020), at the beginning of the experiment, all participants were told that they were going to learn part of a new language called Nápíjò spoken in Southeast Asia. They were told that they would learn the names of some objects and how to describe these objects in this language.

Phases 1–4 In Phases 1-4, participants learned and were tested on the meanings of the novel words through a combination of language-picture matching tasks. They also learned and were tested on describing objects with a noun followed by only one modifier (e.g., *éyày púkù*, *úhù gìmí*).

Phase 5 – NP hierarchy training and similarity judgments: In this phase, participants learned how to describe sets of objects with two postnominal modifiers following either the homomorphic (control group) or the non-homomorphic order (experimental group). A similarity judgment followed each description. Specifically, in each of the 36 trials, participants first saw an image consisting of a set of objects for 2s. Next, a set of buttons containing each of the seven novel words appeared on the screen along with a blank space to be filled in (button location randomized across trials). Participants were asked to click on the correct buttons to describe the set of objects in the image. Crucially, each button choice was sequentially evaluated. At each choice point, the word participants chose was inserted into the blank and a green tick would appear *only* if this word was correct and in the correct order; the blank would not fill and a red cross would appear if the participant clicked the wrong word, or the word was in the wrong order. In other words, at first participants had to guess the correct word order; they had to learn the correct word order for their group by trial-and-error. After all three blanks (one noun followed by two modifiers) were correctly filled, participants proceeded to a similarity judgment. They were asked

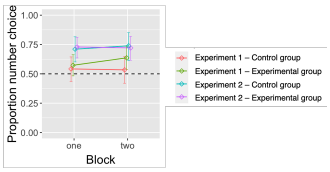

Figure 3: Proportions of number-based similarity judgments in each experiment, group, and block

to select which one of two images was more similar to the target image they just described. One of the two image options differed from the target only in the color of the objects in the set, the other image differed from the target only in the number of objects in the set. To make sure that participants attended to the stimuli adequately, three additional attention check trials were included. In these trials, participants were asked to select which of the two options was exactly the same as the target. See Figure 2 for examples of this phase.

Phase 6 – Debriefing: Finally, participants were asked about what they thought this experiment was about, and what criteria they used to make their similarity judgments.

Results

Six control group and three experimental group participants were excluded for accuracy below 50% in the attention check trials. We thus analysed the data of 40 participants in each group. The critical measure we are interested in is the change in similarity judgment preferences as participants were trained on the order in the language. We divided Phase 5 into two blocks to assess this, as seen in Figure 3. Recall that participants learned both orders gradually, however we expect that, for English speakers, the homomorphic order conforms to their expected hierarchy (given previous experimental results, e.g., Martin et al., 2020). By contrast, the non-homomorphic order does not conform to the expected hierarchy. Instead, this order potentially implies a tighter relationship between N and Num. This should lead to a change in the judgments of experimental group participants who learned the non-homomorphic condition.

We fit mixed logit models using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in R (R Core Team, 2022) to predict the likelihood of judging images that match the target in number as more similar than those that match the target in color, in other words, the likelihood of making similarity judgments based on number (abbreviated as "Judge", similarity judgment based on number is treatment coded as 1; those based on color is treatment coded as 0). The fixed effects are Group (control vs. experimental condition), Block (one vs. two), and their interaction. Following the principle outlined in Barr, Levy, Scheepers, and Tily (2013), throughout this study, the selected model is the one with the maximal random effect structure justified by the design that converges. Here, the selected model is Judge $\sim$ Block*Group + (1 + Block | participant) + (1 | trial) + (0 + Group | trial). This model shows no significant main effect of either Group or Block (both p $>$.05), but reveals a significant interaction between Group and Block (estimate = 0.66, se = 0.30, z = 2.17, p = 0.030). To break down this interaction, we fit models on each group separately with Block as the only fixed effect. The selected model for the control group (Judge $\sim$ Block + (1 | participant)) reveals no significant effect (estimate = -0.05, se = 0.14, z = -0.38, p = 0.704); but the selected model for the experimental group (Judge $\sim$ Block + (1 + Block | participant) + (1 | trial)) shows a significant effect of Block (estimate = 0.61, se = 0.25, z = 2.42, p = 0.015). This suggests that the likelihood of making number-based similarity judgments was significantly higher in Block two than Block one only in the experimental group, i.e., the group who learned the non-homomorphic order implying a tighter relationship between objects and their number compared to objects and their color; the likelihood in the control group did not significantly differ between the two blocks. The increasing preference for color in the experimental group is unlikely a result of participants' conscious strategies, as in the debriefing, no experimental group participants reported any relationship between similarity judgment preferences and word order. [2]

Experiment 2

As discussed in the Introduction ("Thinking for speaking" section), previous studies have found that language effects on cognition may be reduced or even removed completely when online verbal involvement is no longer encouraged, or is actively interfered with. If this also happens in our case, it would strengthen our evidence for a link between language and cognition. Therefore, in Experiment 2 we interfered with online language involvement using a secondary task of digit recall. We also had participants complete language training *before* doing similarity judgments so that language usage was activated less during the critical task.

Participants

Another 75 English native speakers without prior knowledge of non-homomorphic languages who had not participated in Experiment 1 were recruited via Prolific and completed the experiment. They had been randomly assigned to the experimental (N = 39, Mage = 43.85, 27 females) or the control group (N = 36, Mage = 37.92, 16 females).

Stimuli and procedure

The stimuli were identical between the two experiments. In addition, phases 1-4 of Experiment 2 were exactly the same as Experiment 1. After this, the two diverged. In Experiment 1, the fill-in-the-blank task was interspersed with the similarity judgment task. In Experiment 2, we first trained participants on the homomorphic (control group) or the non-homomorphic word order (experimental group) via the fill-

[2] However, two control group participants did report such a relationship. But since the critical observation for Experiment 1 lies in the experimental but not the control group, we still argue that the increasing preference for number along the training on the non-homonorphic order is unlikely to be due to conscious strategies.

in-the-blank task. After completing this training, they completed similarity judgments during online verbal interference. This means that the two experiments differ both in the presence of online verbal interference, and in the interval between the verbal descriptions and similarity judgments. To partially compensate for this, we increased the amount of fill-in-the-blank trials in Experiment 2 (from 39 trials in Experiment 1 to 58 in Experiment 2).

The verbal interference task worked as follows. On each trial, participant first heard a five-digit string that they were asked to remember. Next, they performed the similarity judgment as described for Experiment 1. Afterwards, they typed the five-digit string in a box. We reason that to perform the digit recall task, participants had to rehearse the digits in the phonological loop of their working memory (Baddeley, 1998), thereby interfering with their online recruitment of language during the similarity judgment. Three attention check trials were also included, as in Experiment 1.

Results

Three participants in the control group and one in the experimental group were excluded for digit recall accuracy being below 50%; another two control group participants were excluded for accuracy being below 50% in the attention check trials. This resulted in the analysis of 31 and 38 participants in the control and experimental group respectively. Although here we do not expect to see any effect of block, for ease of comparison with Experiment 1, we also divided Phase 5 of Experiment 2 in halves. See Figure 3 for the proportion of number-based similarity judgments in each group, block, across both experiments. To analyse this data, we again fit mixed logit models. In our initial analysis, we model only Experiment 2, with a model predicting the likelihood of making number-based similarity judgments using Group. The selected model (Judge $\sim$ Group + (1 | participant) + (1 + Group | trial)) showed no significant effect of Group (estimate = -0.01, se = 0.54, z = -0.01, p = 0.989). No participants reported any relationship between similarity judgments and the word order.

To directly compare the two experiments, we analysed the similarity judgment data from both Experiments 1 and 2, including Block.[3] We fit a mixed logit model predicting the likelihood of making number-based similarity judgments using the following fixed effects: Group (control vs. experimental), Block (one vs. two), Experiment (one vs. two), and all their interactions. The selected model (Judge $\sim$ Group*Experiment*Block + (1 | participant) + (1 | trial) + (0 + Experiment | trial)) revealed a significant main effect of Experiment (estimate = 1.28, se = 0.56, z = 2.30, p = 0.022), a significant two-way interaction between Group and Block (estimate = 0.47, se = 0.20, z = 2.42, p = 0.016), and a significant three-way interaction (estimate = -0.81, se = 0.31, z = -2.64, p = 0.008). The main effect of Experiment suggests that the likelihood of making number-based similarity judgments was significantly higher in Experiment 2 than Experiment 1 regardless of Group and Block (to be explored further in the Discussion). To break down the significant three-way interaction, we fit models on each experiment respectively. The model for Experiment 1 has been reported above in the result section for Experiment 1. Two models were selected for Experiment 2, as they exhibited equal complexity of the random effect structure and the same AIC and BIC values: Judge $\sim$ Block*Group + (1 | participant) + (0 + Block | participant) + (1 + Group | trial); Judge $\sim$ Block*Group + (1 + Block | ID) + (1 | trial) + (0 + Group | trial). Neither model revealed any significant effects (all p $>$.05).

[3] Although we don't expect Block to be a significant predictor in Experiment 2, we include it to enable us to ensure that the interaction effect we see in Experiment 1 is reliable in comparison.

Discussion

In this study, we investigated whether "thinking for speaking" effects can be triggered by differences in complex syntax. In particular, we targeted the structure of complex NPs. In Experiment 1, we trained participants on novel lexical items describing unfamiliar objects either in a homomorphic order—i.e., an order which conforms to a purportedly preferred NP hierarchy, in having Adj closer to N than Num—or a non-homomorphic order—i.e., an order which instead implies a hierarchy in which Num is closer to N than Adj. Interspersed with this training, they were asked to provide similarity judgments between the image just described and two others that differed in numerosity or color. We found that training on a non-homomorphic order shifted participants' similarity judgments: they increased their likelihood of basing judgments on shared number across images rather than shared color. This is consistent with the idea that having Num closer to N in the syntax increases the perceptual or cognitive importance of numerosity in a task that actively involves language use (here a fill-in-the-black image description task). In Experiment 2, we found that this effect was reduced to non-significant level when similarity judgments were provided only after order training, and in the presence of verbal interference.

It is worth noting that in Experiment 1, the similarity judgments of the control group did not change significantly over time. This might indicate that the control group's relative preference for color over number was already at ceiling level when the experiment started. Here, ceiling level does not mean choosing color 100% of the time. Instead, it means that, given the perceptual difference between the color and number dimensions in this specific task, there might be no room for the control group's preference towards color to further increase. At the beginning of Experiment 1, both groups' preference towards number was numerically above chance. Once again, this does not mean that the hierarchical knowledge participants come to the task with failed to have an effect on similarity judgments at the start. Rather, it could be that, in this particular task, the number dimension was just more salient than the color dimension. Differences in relative salience of this kind are difficult to control. However, a priori expectations about similarity could be tested using Kîîtharaka

speakers, keeping in mind issues around inferring a linguistic cause in cross-cultural comparison.

The findings of Experiment 1 suggest that the effect of language on cognition during language use can be extended beyond the codability of semantic and grammatical categories (e.g., evidentiality, path and manner of motion events, grammatical gender and aspect) to syntactic hierarchy. In our study, though the semantic categories of color and number were both coded, their different status in homomorphic versus non-homomorphic orders nevertheless impacted participants' perceptions of similarity with respect to these categories. Here we have interpreted this effect as driven not by the linear order itself, but through the syntactic hierarchies that are implied by each order: homomorphic orders straightforwardly map to a purportedly preferred representation in which Adj is closest to (i.e., forms an immediate constituent with) N; non-homomorphic orders imply a different structure: in the case tested here, Num is closer. Describing the target set of objects following the non-homomorphic hierarchy may guide participants to attach a higher weight to number over color, thereby warping subsequent similarity judgments.

It is worth considering, however, whether strategies related to the linear order itself could explain our findings. For example, participants could base their similarity judgments on the modifier that comes first linearly - color for the control group; number for the experimental group. If participants rehearsed verbal descriptions in the phonological loop of their working memory while doing the similarity judgment, the first modifier might be the first cue available for the similarity judgment. However, this might lead us to expect that the two groups' similarity judgments should have already differed in Block 1 of Experiment 1. Moreover, it's not clear that the alternatively—i.e., basing similarity judgments on whatever came *last* linearly—is not more plausible. If the element that came last linearly had the strongest trace in participants' working memory, then we would expect the opposite pattern of results: more number-based responses in the *control* group. Nevertheless, fully ruling out linear strategies would likely require training on pre-nominal modifier order and verifying that e.g., Adj-Num-N and not Num-Adj-N leads to a preference for number-based similarity judgments.

However, our results are in some sense inconsistent with Martin et al. (2024). There, the finding was that speakers of a native language with a non-homomorphic language nevertheless prefer a homomorphic order in an artificial language task. Martin et al. (2024) argue that this supports the claim that the hierarchy is universally active, even when surface order doesn't conform to that. However, the task we use here is very different; it actively measures changes in perceived importance of a conceptual category during controlled training only on complex NP order. It may be that this kind of task is required to detect changes in conceptual representations. It also remains to be seen whether the opposite is possible, i.e., whether conceptual representations can impact syntactic hierarchies. For example, ongoing work is investigating whether exposure to a distribution of objects and properties in which numerosities are more predictive of the objects they modify than adjectival properties (e.g., color, or texture), induces the expectation of non-homomorphic word order.

Notably, in Experiment 2, no effect of language training was found when language use was no longer interspersed with similarity judgment trials, and was indeed actively interfered with using verbal interference. Arguably, this finding strengthens the claim that the effects of language on cognition occur *during* language use. However, future testing is needed to determine whether the lack of effect in our case was due to online verbal interference alone or the prolonged time interval between the verbal description and the similarity judgment. It is worth noting that this timecourse is quite similar to that used in previous "thinking for speaking" studies (e.g., participants describe a set of images, and later provide judgments, Wang & Li, 2023; Park 2020).Nevertheless, we do not intend to draw any strong conclusion based on a direct comparison between the two experiments. Previous studies suggest that it is very difficult, if not impossible, to develop two tasks that only differ in verbal involvement and nothing else (e.g., see Perry & Lupyan, 2013 for discussion). Another intriguing finding is that the preference towards number was significantly higher in Experiment 2 than Experiment 1, regardless of the order participants were trained on. Our guess is that the verbal interference task of digit recall drew participants' attention to number overall.

By design, our study only measured the syntactic hierarchy effect on cognition when language is overtly involved (Experiment 1) or disrupted (Experiment 2). It still remains to be seen whether such an effect would still exist when online language use is neither encouraged nor disrupted. If so, then the scope of linguistic relativity effects (Whorf, 1956) could be extended to syntactic hierarchies. It also remains to be seen whether, through long-term experience with certain language(s), syntactic hierarchies can permanently reorganize speakers' conceptual representations as the salience hypothesis posits (see Papafragou and Selimis, 2010). If this were true, we would expect that a syntactic hierarchy effect on similarity judgments can still persist even under online verbal interference for native speakers of non-homomorphic languages.

To summarise, here we have found evidence that syntactic hierarchies can impact conceptual representations during language use, thereby extending the scope of "thinking for speaking" from the codability of semantic and/or grammatical categories to complex syntax. The present study opens up a number of further avenues of research on the relationship between syntactic and conceptual representations.

Acknowledgments

This project was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757643).

References

Abels, K., & Neeleman, A. (2012). Linear asymmetries and the LCA. *Syntax*, *15*(1), 25–74.

Aristotle. (350BC). *On interpretation*.

Baddeley, A. (1998). Working memory. *Comptes rendus de l'Academie des sciences. Serie III, Sciences de la vie*, *321*(2-3), 167-173.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1-48.

Bylund, E., & Athanasopoulos, P. (2024). Thinking for speaking. In T. Ionin, S. Montrul, & R. Slabakova (Eds.), *The routledge handbook of second language acquisition, morphosyntax, and semantics* (p. 37-46). Routledge.

Casasanto, D. (2008). Who's afraid of the big bad whorf? cross-linguistic differences in temporal language and thought. *Language Learning*, *58*(1), 63-79.

Choi, S., McDonough, L., Bowerman, M., & Mandler, J. (1999). Early sensitivity to language-specific spatial categories in english and korean. *Cognitive Development*, *14*, 241–268.

Cinque, G. (2005). Deriving Greenberg's Universal 20 and its exceptions. *Linguistic Inquiry*, *36*(3), 315-332.

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, *111*(16), 5842–5847.

Culbertson, J., Schouwstra, M., & Kirby, S. (2020). From the world to word order: deriving biases in noun phrase order from statistical properties of the world. *Language*, *96*(3).

Dryer, M. (2018). On the order of demonstrative, numeral, adjective and noun. *Language*, *94*(4), 798–833.

Ferrigno, S., Cheyette, S. J., Piantadosi, S. T., & Cantlon, J. F. (2020). Recursive sequence generation in monkeys, children, us adults, and native amazonians. *Science Advances*, *6*(26), eaaz1002.

Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, *279*(1747), 4522–4531.

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

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219–224.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), *Universals of language* (p. 73-113). Cambridge, MA: MIT Press.

Hickmann, M., Engemann, H., Soroli, E., Hendriks, H., & Vincent, C. (2017). Expressing and categorizing motion in french and english. In I. Ibarretxe-Antunano (Ed.), *Motion and space across languages: theory and applications* (p. 61-94). John Benjamins.

Horst, J. S., & Hout, M. C. (2016). The novel object and unusual name (noun) database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393–1409.

Martin, A., Abels, K., Ratitamkul, T., & Culbertson, J. (2019). Cross-linguistic evidence for cognitive universals in the noun phrase. *Linguistics Vanguard*, *5*(1).

Martin, A., Adger, D., Abels, K., Kanampiu, P., & Culbertson, J. (2024). A universal cognitive bias in word order: Evidence from speakers whose language goes against it. *Psychological science*, *35*(3), 304–311.

Martin, A., Holtz, A., Abels, K., Adger, D., & Culbertson, J. (2020). Experimental evidence for the influence of structure and meaning on linear order in the noun phrase. *Glossa*, *5*(1), 1-21.

Montero-Melis, G., Jaeger, T. F., & Bylund, E. (2016). Thinking is modulated by recent linguistic experience: Second language priming affects perceived event similarity. *Language Learning*, *66*, 636–665.

Papafragou, A., & Selimis, S. (2010). Event categorisation and language: a cross-linguistic study of motion. *Language and Cognitive Processes*, *25*(2), 224-260.

Perry, L. K., & Lupyan, G. (2013). What the online manipulation of linguistic activity can tell us about language and thought. *Frontiers in Behavioral Neuroscience*, *7*, 1-4.

Pinker, S. (1994). *The language instinct: the new science of language and mind*. Penguin.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Samuel, S., Cole, G., & Eacott, M. (2019). Grammatical gender and linguistic relativity: A systematic review. *Psychonomic Bulletin and Review*, *26*, 1767–1786.

Slobin, D. (1996). From "thought and language" to "thinking for speaking". In J. Gumperz & S. Levinson (Eds.), *Rethinking linguistic relativity* (chap. 3). Cambridge University Press.

Slobin, D. (2004). The many ways to search for a frog: Linguistic typology and the expression of motion events. In S. Strömqvist & L. Verhoeven (Eds.), *Relating events in narrative: Vol. 2. typological and contextual perspectives* (chap. 10). Mahwah, NJ: Lawrence Erlbaum Associates.

Soroli, E. (2024). How language influences spatial thinking, categorization of motion events, and gaze behavior: a cross-linguistic comparison. *Language and Cognition*, *16*(4), 924-968.

Talmy, L. (2000). *Toward a cognitive semantics: Vol. ii: Typology and process in concept structuring*. MIT Press.

Tomasello, M. (2003). Constructing a language: A usage-based theory of-language acquisition: Harvard. *Press, Cambridge, MA and London*.

Vanek, N. (2020). Changing event categorization in second language users through perceptual learning. *Language Learning*, *70*(2), 309-348.

Vanek, N., & Selinker, L. (2017). Covariation between temporal interlanguage features and nonverbal event categorisation. *International Review of Applied Linguistics in Language Teaching*, *55*, 223-243.

Wang, Y., & Li, W. (2023). Multilingual learning and cognitive restructuring: The role of audiovisual media exposure in cantonese–english–japanese multilinguals' motion event cognition. *International Journal of Bilingualism*, *27*(3), 331-348.

Whorf, B. L. (1956). *Language, thought, and reality; selected writings*. MIT Press.

Ünal, E., & Papafragou, A. (2020). Relations between language and cognition: Evidentiality and sources of knowledge. *Topics in Cognitive Science*, *12*, 115-135.