Investigating event orders in LLMs: Insights from English, Norwegian, and Greek temporal connectives

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Rognan, Hannah S.

Grigoroglou, Myrto

Bhattasali, Shohini

Publication Date

2026

Investigating event orders in LLMs: Insights from English, Norwegian, and Greek temporal connectives

Hannah S. Rognan[1,2] **Myrto Grigoroglou**[1] **Shohini Bhattasali**[1,3]
hannah.rognan@uantwerpen.be **m.grigoroglou@utoronto.ca** **shohini.bhattasali@utoronto.ca**

[1]Department of Linguistics, University of Toronto
[2]CLiPS, University of Antwerp
[3]Department of Language Studies, University of Toronto Scarborough

Abstract

Event knowledge representations in Large Language Models (LLMs) are explored through the lens of temporal connectives. Temporal connectives like *before* and *after* play an important role in sentence comprehension by linking events in the sentence to real-world event order. Their position in a sentence can create different outcomes for event order interpretation and sentence comprehension, and they have therefore been studied extensively in experimental research. Importantly, previous studies have found a human preference for chronological ordering in sentence comprehension. The current study investigates whether LLMs can recognize event order and reflect the human-like processing patterns of temporal connectives across three languages: English, Norwegian and Greek. The aim of the study is to shed light on how sentence structure and event sequence influence LLM predictions, with implications for both cognitive modeling and the knowledge representations learned by LLMs. Results vary by language and show that some models do have event representations and reflect human-like patterns of temporal ordering. The results highlight that language models for some languages (English) are better optimized for cognitive modeling than others (Greek) and underscore the need for more cognitively motivated evaluation benchmarks to assess the models being used in cognitive science research.

Keywords: temporal connectives; event order; large language models; cognitive modeling

Introduction

The linguistic and cognitive abilities of Large Language Models (LLMs) are increasingly being investigated, particularly with respect to the kinds of knowledge they encode. LLMs are powerful neural network models that learn how to predict words (given a context), and capture the relationships between them, through training on large amounts of textual data. Through the data and the training process, in which it learns how to make the best predictions, the model comes up with the likelihood of words based on contexts. This probability distribution allows for the study of expectancy effects in models, making them useful for cognitive modeling. Unlike older sequential language models, this probability distribution is not solely frequency-driven since the LLMs are able to capture more complex relationships due to their self-attention mechanism, which allows them to weigh the importance of different input words simultaneously and capture long-range dependencies in natural language (Vaswani et al., 2017). Given these advanced capabilities of LLMs, it is an open question whether the models that support their ability to produce coherent language lead to an ”understanding” of the conceptual structures underlying language.

To answer this question, contemporary computational linguistics research has focused on how well LLMs compare to humans in linguistic processing by investigating the performance of LLMs on adapted experimental tasks that have previously been used in psycholinguistic experiments with human participants. Prior research has illustrated how this approach can help us examine whether LLMs and humans might share knowledge representations (such as syntactic structure, pragmatic inferences, etc.) that support natural language comprehension (Futrell et al., 2019; Ettinger, 2020; Pandia, Cong, & Ettinger, 2021; Schrimpf et al., 2021; Davis & van Schijndel, 2021; Mahowald et al., 2024; Tuckute, Kanwisher, & Fedorenko, 2024).

In this work, we investigate whether LLMs have knowledge of event representations, by focusing on sentences with temporal connectives. Temporal connectives require language users to map real-world events (that happened at different time points) onto syntactic structure and can therefore serve as a test for understanding how events occurred in the real world. The connectives *before* and *after* are an ideal testbed for examining LLMs' understanding of event order, as their syntactic position in a sentence can lead to different event order interpretations. Specifically, when the connective *after* appears in sentence-initial position as in (1), the sentence reflects the chronological order that the events occurred, while when the connective *after* appears in mid-sentence position as in (2), the sentence reflects the reverse chronological order. Conversely, when the connective *before* appears in sentence-initial position as in (4), the sentence reflects reverse chronological order, but when it appears mid-sentence as in (3), events are interpreted chronologically [1].

(1) **After** he blew out the candles, the boy cut the cake.
CHRONOLOGICAL

(2) The boy cut the cake **after** he blew out the candles.
REVERSE

(3) The girl poured the syrup **before** she ate the pancakes.
CHRONOLOGICAL

(4) **Before** she ate the pancakes, the girl poured the syrup.
REVERSE

[1]Example sentences from Keller-Cohen (1987)

Temporal connectives have been studied extensively in experimental research with both adults (Politzer-Ahles, Xiang, & Almeida, 2017; Cain, Patson, & Andrews, 2005; Nieuwland, 2015), and children (Blything & Cain, 2016; Blything, Davies, & Cain, 2015; de Ruiter, Theakston, Brandt, & Lieven, 2018; de Ruiter, Lieven, Brandt, & Theakston, 2020; Feagans, 1980; French & Brown, 1977; Stevenson & Pollitt, 1987; Trosborg, 1982). This research shows that sentences reflecting chronological order are in general easier to process than sentences reflecting reverse order across different experimental paradigms, age groups and languages.

However, other factors such as the ordering of main and subordinate clause or the givenness of the information expressed in the complex sentence have also been shown to affect the interpretation of sentences with temporal connectives (de Ruiter et al., 2020; Makrodimitris & Schulz, 2025; Pyykkönen & Järvikivi, 2012). Given that languages vary greatly in terms of word order, it is an open question how these different factors affect interpretation of sentences with temporal connectives. The present study aims to address this gap by looking at data from three languages with varying word orders (English, Norwegian, and Greek), and utilizing LLMs as proxy psycholinguistic subjects (Futrell et al., 2019), to investigate whether the models are sensitive to event order and whether they show the chronological preference across these languages.

Current Study

We examine temporal connectives across three languages by conducting masked language modeling experiments. These computational experiments are based on prior psycholinguistic experiments with human participants. In addition to the ambiguous sentences used in the previous study (Pandia et al., 2021), we extend the type of sentences we use to unambiguous sentences, and use model prediction scores as an evaluation metric, to gain further insight into whether LLMs are sensitive to event orders in sentences (e.g., chronological vs. reverse) and whether they have any event knowledge representations. Through a series of three experiments in English, Norwegian, and Greek, we aim to answer the following questions in the current study:

1. Can LLMs detect event orders in sentences?
2. Do LLMs make predictions in ways that mirror human cognition?

Cross-linguistic Approach

English is the most widely-studied language both in the psycholinguistic research on temporal connectives (Politzer-Ahles et al., 2017; Blything et al., 2015) and in comparisons between the linguistic abilities of LLMs versus humans (Pandia et al., 2021; Futrell et al., 2019). We extend this line of work cross-linguistically to incorporate two understudied languages in computational modeling, Norwegian and Greek, in order to distinguish between competing explanations for event order prediction effects in language models. The temporal connectives we investigate in Norwegian are *før* ’before’ and *etter* ’after’ and πριν/*prin* ’before’ and αφού/*afu* ’after’ in Greek.

In addition to availability of language models and data, English, Norwegian, and Greek were chosen for comparison because of word order differences. Although all three share SVO (subject-verb-object) order, English maintains this order strictly, Norwegian permits reordering under verb-second (V2) constraints, and Greek allows considerable word order flexibility due to extensive morphological marking. These differences are particularly relevant for sentences involving temporal connectives, as changes in the position of the temporal connective may or may not preserve surface word order across languages.

For example, in English, changing the position of a temporal connective does not alter the ordering of words inside the two clauses (*She drinks coffee after breakfast* vs. *After breakfast she drinks coffee*). In Norwegian, however, moving the connective from the medial to the initial position causes the verb and the subject to switch positions in the main clause (5 vs. 6 below). Greek, on the other hand, allows multiple orders of the subject and verb while preserving the same temporal interpretation. Thus, unlike Norwegian, placing the temporal connective at the beginning of the sentence does not force a specific reordering in Greek; rather, different surface orders remain grammatical and semantically equivalent. This flexibility is illustrated in (7) and (8), where the clause-internal word order remains the same despite the change in connective position.

(5) *Hun drikker kaffe etter frokost*
She drinks coffee after breakfast
’She drinks coffee after breakfast’

(6) *Etter frokost drikker hun kaffe*
After breakfast drinks she coffee
’After breakfast she drinks coffee’

(7) *Pini kafe to koritsi afu fai proino.*
Drinks coffee the girl, after she-eats breakfast.
’She drinks coffee after breakfast’

(8) *Afu fai proino, to koritsi pini kafe.*
After she-eats breakfast, the girl drinks coffee.
‘After breakfast the girl drinks coffee.’

By testing LLMs across languages with distinct word order constraints, we can assess whether observed preference patterns are attributable to structural properties of sentences or to more abstract representations of event order. If models exhibit a consistent preference for chronological ordering across all three languages, this would suggest that the effect cannot be reduced to language-specific word order patterns and instead reflects learned representations of temporal coherence. Conversely, if the effect is restricted for instance to English, where word order remains constant across connective positions, this would support the interpretation that models rely primarily on surface co-occurrence patterns rather than event representations.

Models & Experimental Design

Across the three languages, we used BERT (Bidirectional Encoder Representations from Transformers) models, which are Transformer-based language models that learn deep bidirectional representations by pretraining on masked language modeling and next sentence prediction tasks. Unlike decoder models, which generate predictions based only on preceding contexts, encoder models such as BERT have access to both preceding and following contexts, enabling predictions to be informed by the relation between the two events in each sentence, and aligning our design with cloze-style tasks in psycholinguistics.

We used the English model $\text{BERT}_{\text{BASE}}$ (Devlin, Chang, Lee, & Toutanova, 2019), the Norwegian model $\text{NorBERT}_{\text{BASE}}$ (Samuel et al., 2023), and the Greek model $\text{GREEK-BERT}_{\text{BASE}}$ (Koutsikakis, Chalkidis, Malakasiotis, & Androutsopoulos, 2020). While multiple BERT-style models are available for English, we chose $\text{BERT}_{\text{BASE}}$, which is relatively small, in order to be comparable to the size of $\text{NorBERT}_{\text{BASE}}$ and $\text{GREEK-BERT}_{\text{BASE}}$, and as such ensured our experiments were comparable across languages.

LLMs can be used to conduct masked modeling experiments because they are trained to predict a masked word given the rest of the words in the sentences. Masked language modeling experiments, which we use to infer temporal connective preferences in the current study, is setup in a similar way: the trained model is provided with a sentence where one word is hidden (with *[MASK]*) and the tokens that the model would predict for the masked word, along with the prediction score, are extracted. To make a prediction about the hidden masked word, the model uses the context that surrounds the hidden word, including lexical, sequential and grammatical cues. The predicted tokens and their prediction scores can thus be used to understand which words the model finds the most likely in a given sentence. For example, given the masked sentence: *The boy cut the cake **[MASK]** he blew out the candles.*, the LLM could make the following predictions:

{*'score': 0.7, 'sequence': The boy cut the cake **after** he blew out the candles.*}
{*'score': 0.2, 'sequence': The boy cut the cake **and** he blew out the candles.*}
{*'score': 0.1, 'sequence': The boy cut the cake **before** he blew out the candles.*}

Model preferences of event order were determined through prediction scores, as illustrated above. For each sentence, there is a [MASK] in the position where the connective would be, and we considered the prediction scores that the model assigned to each of *before* and *after* (or their translation equivalents) in the masked position. To do so, we extracted the top 50 tokens that the model predicts for the masked word, and then looked at the score for each of the two connectives. The connective that received the highest prediction score out of the two temporal connectives under consideration, was the connective that we considered to be the model's preference for the sentence. Since we know temporal connectives can appear in different positions in the sentence, we have two conditions: **sentence-initial** and **sentence-medial**. In the sentence-initial condition, the hidden, masked word is at the beginning of the sentence. In the sentence-medial condition, the hidden, masked word is in the middle of the sentence.

Experiment 1: Chronological vs. reverse in ambiguous sentences

Objective

The aim of this experiment was to see which of the two temporal connectives the model preferred when given sentences with ambiguous order of events (i.e., two events without a logical temporal relationship between them). The choice of temporal connective would reflect the model's preference in terms of chronological vs. reverse order.

Data

We tested 155 ambiguous sentences from prior experimental stimuli (Politzer-Ahles et al., 2017). Each sentence in our data has a version where the masked word is placed sentence-initially as well as sentence-medially, and the masked word (the connective) can occur in either of these positions. The same sentences were used across the three languages. The Norwegian sentences were first translated from English into Norwegian by one of the authors, a native speaker, and then later verified by two additional Norwegian linguistic consultants. The Greek sentences were first translated from English to Greek through the Google Translate API, and then later reviewed and verified by native Greek speakers through Prolific (Prolific, 2025). An example of an ambiguous sentence is shown in Table 1.

Condition	Example item
Sentence-Initial	***[MASK]*** the diver reached the wreck, the fishing boat cast its nets.
Sentence-Medial	The fishing boat cast its nets ***[MASK]*** the diver reached the wreck.

Table 1: Ambiguous example sentence used for masked modeling experiments in sentence-initial and sentence-medial conditions.

Predictions

Following the chronological ordering preference observed cross-linguistically, we expected *after* to be preferred by the model in sentence-initial condition and *before* to be preferred by the model in sentence-medial condition. Specifically, if the prediction score for *before* was higher in the sentence-medial condition and for *after*, it was higher in the sentence-initial condition, this would illustrate the chronological preference of the LLMs and mirror human-like patterns. If it was the other way around, it would indicate that the LLMs preferred the reverse order.

Results

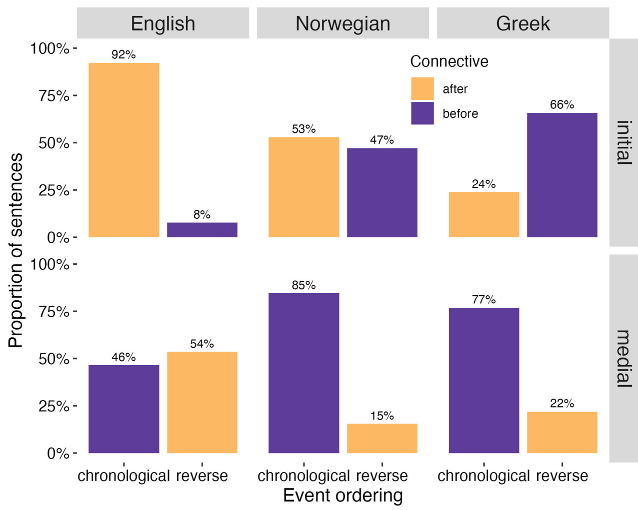

Figure 1: Model preferences for ambiguous sentences.

Overall, we corroborated the earlier pattern found for English where the chronological preference was only observed in the sentence-initial condition. The Norwegian results patterned similarly to English but for Greek, we found a preference for the reverse order.

English: Results in the sentence-initial condition (Fig. 1) shows that, in the initial condition, there was a clear preference for the connective that produced the chronological order of events (*after*), but in the medial position of the masked word there was no clear preference for the chronological event order over the reverse event order, which is consistent with earlier findings (Pandia et al., 2021).

Norwegian: There was a clear chronological preference pattern only in one of the two positions for the Norwegian model. Interestingly, however, the pattern was exactly the opposite of the English experiment, as seen in the Norwegian sentence-medial condition (Fig. 1). While the English model had a clear preference for the chronological connective only in the initial condition, the Norwegian model had this preference only in the medial condition, with 85% of the sentences having the chronological connective (*før* 'before') as the most likely of the two connectives. On the other hand, in the initial condition, the model chose the chronological connective only 53% of the time, and thus the pattern did not show up in that position.

Greek: We did not see the same clear pattern of chronological order preference, when looking at the preferences of the Greek model with the ambiguous sentences, as seen in Figure 1. In the first two models, there was a clear chronological preference in only one of the two positions, and in the second position, the predictions were almost equal between the chronological and reverse order. However, with the Greek model, we see that the model prefers the chronological connective in the medial condition, but in the initial condition there is a preference for the reverse condition. As evidenced by the similar proportion of sentences that the model predicted *prin* 'before' for in both conditions, it looks like the model generally prefers to predict *prin* 'before', rather than a specific ordering of events.

Experiment 2: Chronological vs. reverse in unambiguous sentences

Objective

The aim of this experiment was to see which of the two temporal connectives the model preferred when given sentences with unambiguous order of events (i.e., two events that do have a logical temporal relationship between them). The choice of temporal connective would reflect the model's preference in terms of chronological vs. reverse order. Earlier studies with LLMs only looked at ambiguous sentences (Pandia et al., 2021), and that could be one of the factors explaining why a clear chronological preference was not observed across the sentence-initial and sentence-medial conditions.

Data

We tested 64 unambiguous sentences; 32 sentences are experimental stimuli from two prior psycholinguistic studies, Cain and Nash (2011) and Keller-Cohen (1987), and we created 32 sentences on our own to supplement these sentences. Similar to Experiment 1, each sentence in our data has a version where the masked word is placed sentence-initially as well as sentence-medially, and the masked word (the connective) can occur in either of these positions. The same sentences were used across the three languages and we followed the same methods, as described in an earlier section, to translate from English to Norwegian and Greek and verify the translations. Example unambiguous sentences are shown in Table 2.

Conditions	Target	Example item
Sentence-Initial	*before*	***[MASK]*** he went to sleep, Jack watched TV.
Sentence-Medial	*after*	They got on the bus ***[MASK]*** it arrived.

Table 2: Unambiguous example sentences used for masked modeling experiments in sentence-initial and sentence-medial conditions.

Predictions

Following the chronological ordering preference observed cross-linguistically, we expect *after* to be preferred by the model in sentence-initial positions and *before* to be preferred by the model in sentence-medial positions. We introduced unambiguous sentences in Experiment 2 because we wanted

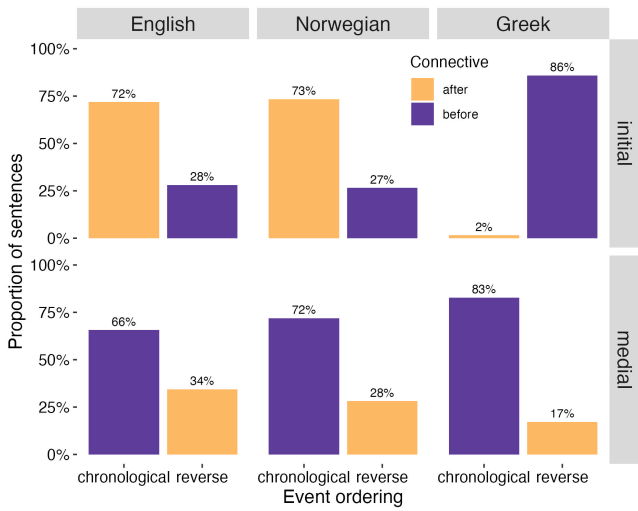

Figure 2: Model preferences for unambiguous sentences.

to examine when there is a clear temporal relationship present between the events in the sentence, if the model would have a stronger preference in both conditions. As we introduced sentences with an unambiguous, logical temporal relationship between events, we expected results to be different than in the first experiment. Specifically, we predicted that, for all languages, we would see a more clear preference for the chronological ordering (preference for *before* in sentence-medial; *after* in sentence-initial), which would indicate that the models' are sensitive to event ordering and thus, this would provide evidence that they have event knowledge representations.

Results

Overall, we did see a human-like processing pattern of chronological event order preference when unambiguous sentences were introduced in Experiment 2.

English: Fig. 2 shows that the English model predicted the connective that produced a chronological order of events both in the initial position and in the medial position with the unambiguous sentences, in contrast the results from Experiment 1 with the ambiguous sentences where the pattern was only present for the initial position.

Norwegian: The model preferred the connective that would produce the chronological order in both positions, with over 70% sentences having the chronological connective as the connective with the highest prediction score. This pattern differs from the results from Experiment 1, where the pattern of preferring the chronological connective was only found in the medial position, thus demonstrating that using unambiguous sentences also seems to affect predictions in the Norwegian model.

Greek: The model did not produce the same pattern of results as the first two languages. The results show a preference for reverse order (*prin* 'before') in the sentence-initial position and chronological order (*prin* 'before') in the sentence-medial position. Therefore, similarly to the results from the first experiment, there was a clear preference in the Greek model for the *prin* 'before' connective in Experiment 2 and not a clear preference between chronological vs. reverse event order.

Experiment 3: Model accuracy for temporal connective prediction

Objective

While the first two experiments were designed to examine whether the models would reflect a human-like cross-linguistic preference for chronological ordering, this experiment examines whether the model can predict the correct temporal connective, regardless of chronological vs. reverse order.

Data

We tested the same unambiguous sentences from Experiment 2. Only unambiguous sentences were used because there is only one of the two connectives that would be reasonable to humans in these sentences, and we measured whether the model accurately predicted this target connective. Following Experiment 1 & 2, we tested sentences from all three languages.

Predictions

In this experiment, we wanted to examine whether the model could accurately predict the logical order of events in a sentence, using event representations that define the natural temporal relationship between events in the real world. There are cases where the reverse order is logically more accurate (e.g., in *The flight took off [MASK] they boarded.*, the connective *after*, which reflects the reverse order, is accurate). So, we predict that in these unambiguous sentences the model would show an accurate ordering of events and choose the connective that would produce logical ordering, regardless of sentence-initial vs. sentence-medial conditions. This would provide evidence for event knowledge representations in the models as the results would illustrate they are sensitive to logical ordering of events.

While accuracy around 0.5 is interpreted as chance, in the reverse event order sentences, any non-zero accuracy would reflect that the model is sensitive to event-order and it is not simply generalizing and predicting the connective that is consistent with the chronological event order (i.e, *before* in the sentence-medial condition and *after* in the sentence-initial condition).

Results

Overall, for both the sentence-initial and sentence-final conditions, we observed that the model predicted the accurate connective for English and Norwegian. In addition, when the correct connective was the connective that produced the chronological order, the accuracy was boosted across all languages, except for the sentence-initial condition in Greek.

English: For both the sentence-initial and sentence-medial condition, the accuracy is at 97% when the model predicts

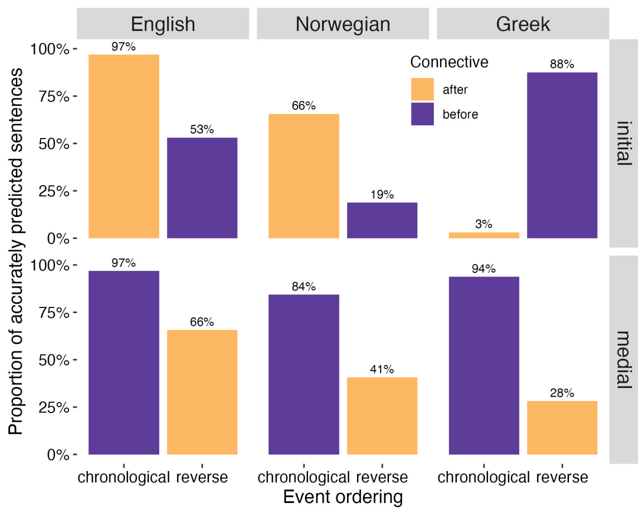

Figure 3: Model accuracy for unambiguous sentences.

the connective that is consistent with the chronological order. For the reverse order, it is at 53% and 66% in the sentence-initial and sentence-medial condition, as seen in Fig. 3.

Norwegian: Similar to English, accuracy is high for the chronological order and lower for the reverse order for both the sentence-initial and sentence-medial conditions.

Greek: The model seems to be unable to predict the connective *afu* 'after', even when it is the connective consistent with the choronological order in the sentence-initial condition and with the reverse order in the sentence-medial condition. The same preference for *prin* 'before' is observed here.

Conclusion

In line with recent studies of LLMs and cognitive modeling, the present study examined whether language models align with patterns of human language behavior. We used temporal connectives as a case study to understand models' representations of event order across languages. The results show that models can detect event orders and they do mirror the human-like preference for chronological ordering. This suggests that the models do have some event knowledge representations, especially when sentences have events with a clear, logical temporal relationship between them.

However, this result varied by language. The current study found that the chronological order preference pattern that has been shown with human subjects is reflected in LLMs across English and Norwegian models, but not in the Greek model. In addition, the LLM's accuracy was high for predicting the correct temporal connective in English, but slightly lower for Norwegian and considerably lower for Greek. Collectively, this study validates the robust utility of LLMs in cognitive modeling, providing evidence that they can develop event knowledge representations and reflect human-like patterns of temporal ordering, mirroring the complex processes observed in human sentence comprehension.

Discussion

A main take-away from the differences between the models in these experiments is that the training data used in models has a considerable impact on the model predictions. Post-hoc analyses revealed that tokens that were of interest (i.e., *prin* 'before' and *afu* 'after') were not highly frequent in the Greek training data, and this could have produced results considerably different in the Greek model than the other two models. This finding is meaningful for the current investigation because the experiments suggest that we need to carefully consider how we are using LLMs in cognitive science experiments. Although some models work well and align with human psycholinguistic experiments, as seen with the English and Norwegian models, the Greek model evidently is not as useful for modeling temporal connective processing. [2]

In addition, cross-linguistic differences in the semantics of the connectives may have also affected the results. For instance, the Greek temporal connective *afu* 'after' can be used both as a temporal and causal connective, and this ambiguity could have been a challenge for the model. Interestingly, this ambiguity has been noted as a particular learning challenge in the acquisition of Greek, leading to later acquisition of *afu* 'after' compared to *prin* 'before' (Tsakali & Vamvouka, 2021). By contrast, the connective *prin* 'before' can appear both as a temporal connective and adverb, which might make its context easier to learn for an LLM (Makrodimitris & Schulz, 2025).

LLMs often have benchmarks, which are a set of standardized tests to indicate how well the model does on certain tasks (GLUE for natural language processing (Wang et al., 2018), SAP for syntactic processing (Huang et al., 2024)). Tasks like temporal connective preferences could lead to more cognitive modeling-motivated benchmarks, which can help us assess whether a particular LLM is suitable for a psycholinguistic research investigation.

As LLMs are trained on large amounts of data, the results could be interpreted as the result of co-occurrence patterns and word frequencies in the training data. However, we rule out frequency-effects as a main explanation for the results through several measures of word-frequency in the data. Frequencies of the two connectives in each language were investigated to rule out single-word frequency effects in the model predictions. A statistical n-gram language model trained on part of the English BERT model's training data was also used to check the likelihood of each connective in the masked position in the sentences from our dataset. None of these measures yielded results that aligned with the predictions of the three models in our experiments, and the experimental results can therefore not be attributed to frequency-related effects.

[2] The current work would also benefit from verification with human participants on the same experiments, especially the Greek sentences which did not produce the same results as the English and Norwegian sentences, and this is left for future work.

Acknowledgments

We acknowledge the support of the UTSC Department Research Fund and the Connaught Fund. We would also like to thank Dave Kush, members of CSoL, and our language consultants.

References

Blything, L. P., & Cain, K. (2016). Children's processing and comprehension of complex sentences containing temporal connectives: The influence of memory on the time course of accurate responses. *Developmental Psychology*, *52*(10), 1517–1529.

Blything, L. P., Davies, R., & Cain, K. (2015). Young children's comprehension of temporal relations in complex sentences: The influence of memory on performance. *Child Development*, *86*(6), 1922-1934.

Cain, K., & Nash, H. M. (2011). The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology*, *103*(2), 429–441.

Cain, K., Patson, N., & Andrews, L. (2005). Age- and ability-related differences in young readers' use of conjunctions. *Journal of Child Language*, *32*(4), 877–892.

Davis, F., & van Schijndel, M. (2021). Uncovering Constraint-Based Behavior in Neural Models via Targeted Fine-Tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 1159–1171).

de Ruiter, L. E., Lieven, E. V., Brandt, S., & Theakston, A. L. (2020). Interactions between givenness and clause order in children's processing of complex sentences. *Cognition*, *198*, 104130.

de Ruiter, L. E., Theakston, A. L., Brandt, S., & Lieven, E. V. M. (2018). Iconicity affects children's comprehension of complex sentences: The role of semantics, clause order, input and individual differences. *Cognition*, *171*, 202–224.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186).

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48.

Feagans, L. (1980). Children's understanding of some temporal terms denoting order, duration, and simultaneity [Journal Article]. *Journal of Psycholinguistic Research*, *9*(1), 41-57.

French, L. A., & Brown, A. L. (1977). Comprehension of before and after in logical and arbitrary sequences. *Journal of Child Language*, *4*(2), 247–256.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 32–42).

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, 104510.

Keller-Cohen, D. (1987). Context and strategy in acquiring temporal connectives. *Journal of Psycholinguistic Research*, *16*(2), 165–183.

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., & Androutsopoulos, I. (2020). Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence* (p. 110–117).

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540.

Makrodimitris, C., & Schulz, P. (2025). Comprehension of complex sentences containing temporal connectives: How children are led down the event-semantic kindergarten-path. *Journal of Child Language*, *52*(3), 615–647.

Nieuwland, M. S. (2015). The truth before and after: Brain potentials reveal automatic activation of event knowledge during sentence comprehension. *Journal of Cognitive Neuroscience*, *27*(11), 2215–2228.

Pandia, L., Cong, Y., & Ettinger, A. (2021). Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Proceedings of the 25th Conference on Computational Natural Language Learning* (pp. 367–379).

Politzer-Ahles, S., Xiang, M., & Almeida, D. (2017). "before" and "after": Investigating the relationship between temporal connectives and chronological ordering using event-related potentials. *PLOS ONE*, *12*(4), e0175199.

Prolific. (2025). *Prolific [online participant recruitment platform]*. https://www.prolific.com. London, UK.

Pyykkönen, P., & Järvikivi, J. (2012). Children and situation models of multiple events. *Developmental Psychology*, *48*(2), 521–529.

Samuel, D., Kutuzov, A., Touileb, S., Velldal, E., Øvrelid, L., Rønningstad, E., ... Palatkina, A. (2023). NorBench – a benchmark for Norwegian language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics* (pp. 618–633).

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Stevenson, R. J., & Pollitt, C. (1987). The acquisition of tem-

poral terms. *Journal of Child Language*, *14*(3), 533–545.
Trosborg, A. (1982). Children's comprehension of 'before' and 'after' reinvestigated. *Journal of Child Language*, *9*(2), 381–402.
Tsakali, V., & Vamvouka, K. (2021). The development of *before* and *after* in greek. In *Proceedings of the 14th International Conference on Greek Linguistics* (pp. 1319–1332).
Tuckute, G., Kanwisher, N., & Fedorenko, E. (2024). Language in brains, minds, and machines. *Annual Review of Neuroscience*, *47*(1), 277–301.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.
Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 353–355).