Modeling the Influence of Verb Aspect on the Activation of Typical Event Locations with BERT

Won Ik Cho
Seoul National University
tsatsuki@snu.ac.kr

Emmanuele Chersoni
The Hong Kong Polytechnic University
emmanuelechersoni@gmail.com

Yu-Yin Hsu
The Hong Kong Polytechnic University
yu-yin.hsu@polyu.edu.hk

Chu-Ren Huang
The Hong Kong Polytechnic University
churen.huang@polyu.edu.hk

Abstract

Prior studies on event knowledge in sentence comprehension have shown that the aspect of the main verb plays an important role in the processing of non-core semantic roles, such as locations: when the aspect of the main verb is imperfective, locations become more salient in the mental representation of the event and are easier for human comprehenders to process.

In our study, we tested the popular language model BERT on two datasets derived from experimental studies to determine whether BERT's predictions of prototypical event locations were also influenced by aspect. We found that, although BERT efficiently modelled the typicality of locations, it did so independently of the verb aspect. Even when the transformer was forced to focus on the verb phrase by masking the context words in the sentence, the typicality predictions were still accurate; in addition, we found aspect to have a stronger influence on the scores, with locations in the imperfective setting being associated with lower surprisal values.

1 Introduction

It has been generally acknowledged in sentence processing research that humans activate generalized event knowledge in the process of understanding natural language sentences (McRae and Matsuki, 2009). Reading/listening verbs (e.g., *open*) activate expectations about their typical arguments (e.g., *door*) (McRae et al., 1998; Ferretti et al., 2001) and vice versa (McRae et al., 2005); the same effect has been found for nouns and their typical co-arguments (Hare et al., 2009). These expectations concerning typical arguments are encoded in the mental lexicon, and are exploited by humans to evaluate the plausibility of verb-argument combinations. Such knowledge is used during sentence processing to generate predictions about upcoming arguments: using different experimental paradigms (e.g. EEG, eye-tracking etc.), previous studies provided evidence that sentences including typical argument combinations are easier for humans to process (Bicknell et al., 2010; Matsuki et al., 2011).

Other studies have investigated the role played by *verb aspect* in event knowledge activation, particularly for non-core roles such as *locations* and *instruments*. Aspect is a grammatical device that denotes the duration, onset, and completion status of an event. According to linguistic theory, a fundamental opposition exists between the *imperfective* aspect (e.g., *The customer was eating in the restaurant*), which describes the event as on-going, and the *perfective* aspect (e.g., *The customer had eaten in the restaurant*), which describes the event as a closed unit and focuses on the resulting state (Madden and Zwaan, 2003).

Based on the above-mentioned literature, Ferretti et al. (2007) used stimulus-onset asynchrony priming and the EEG paradigms to show that, in English, specific expectations were activated for event locations by the verbs describing those events, but only when the verbs were in the imperfective form. Similar findings have been reported for instruments by Truitt and Zwaan (1997) and, more recently, by Madden-Lombardi et al. (2017) in a self-paced reading experiment in French. In line with these findings, Coll-Florit and Gennari (2011) found that imperfective verbs were related to a wider range of semantic associations than perfective ones, because the mental representation of an event as on-going allows one to focus more easily on all the entities that are relevant to the described action (instruments, places, objects, etc.).

In this study, we modelled two datasets on locations using BERT, a state-of-the-art language model (Devlin et al., 2019), and we tested whether verb aspect influenced the model's predictions for upcoming event's locations, by comparing them

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2922–2929

in terms of perfective and imperfective sentences. We found that i) BERT was able to accurately identify typical locations for an event, and that it did so independently of the aspect of the main verb, since the activation levels of locations did not differ significantly across conditions; ii) even when the transformer was forced to focus on the main verb and the other words in the sentence were masked, BERT's typicality predictions were still accurate, and an additional effect of aspect appeared, with locations having significantly lower surprisal values in the imperfective condition.

2 Related Work

Transformer models have become increasingly popular in NLP in recent years (Vaswani et al., 2017; Devlin et al., 2019). The most successful model is probably BERT, which is trained on a *masked language modeling* objective: given the left and the right context of a masked word in a natural language sentence, the model has to predict the word. This conceptually simple yet powerful mechanism has made BERT a very appealing option for NLP researchers working on supervised tasks, and its contextualized representations have taken state-of-the-art performances to new heights.

A number of psycholinguistic-inspired studies designed tests to investigate the actual linguistic abilities of neural network models, including Transformer models. Most of these studies have focused on syntactic phenomena, such as verb-subject agreement and filler-gap dependencies (Linzen et al., 2016; Wilcox et al., 2018; Gulordava et al., 2018; Futrell et al., 2019; Prasad et al., 2019). By contrast, Ettinger (2020) focused on the semantic and pragmatic abilities of the BERT language model by using stimuli from the N400 experiments conducted by Kutas and Hillyard (1984), and showed that the model was strong in associating nouns with their hypernyms, but struggled to handle negations. Close to the spirit of our contribution, Misra et al. (2020) investigated BERT's predictions in a setting aimed at reproducing human semantic priming; they reported that BERT was indeed sensitive to "priming" and predicted a word with higher probability when the context included a related word as opposed to an unrelated one, but this effect decreased in the presence of strongly informative and constraining contexts.

Recent work by Metheniti et al. (2020) has explored the capacity of BERT to reproduce the selectional preferences for verbs - which, from our perspective, was equivalent to modeling the thematic fit of typical event participants (Sayeed et al., 2016; Santus et al., 2017; Chersoni et al., 2020; Marton and Sayeed, 2021). Metheniti and colleagues reported that the correlation of the predictions with human judgements increased when they applied attention masks to the context words in the sentence and forced the model to focus only on the verbs. Finally, Transformers have been used to model typicality effects in language by Misra et al. (2021), although in a different context; i.e. the influence of typicality on category membership judgements.

To the best of our knowledge, the current study is the first to attempt to model argument typicality predictions with BERT for a non-core role (location), and the first to investigate whether and how such predictions are influenced by verb aspect, as in the case in human language processing.

3 Experiments

3.1 Datasets

In our work, we used two datasets that we obtained from the previous studies. The first dataset, which we refer to as **Ferretti07**, consists of the experimental items used by Ferretti et al. (2007) in their EEG experiment. The authors made available a subset of 38 items in which the verb phrase was specifically biased to be followed by a locative prepositional phrase in sentence completion tasks. We excluded 6 of them, in which the location noun was extremely rare and was not included in BERT's basic vocabulary[1]. Each item consisted of an intransitive sentence, with the verb phrase in either the past perfect tense (perfective condition, $PERF$) or the progressive past tense (imperfective condition, $IMPERF$), and a location argument, either typical (TYP) or less typical but still plausible (NON_TYP) (see Example 1).

(1) a. The boy had fished at the lake. ($PERF, TYP$)
b. The boy was fishing at the lake. ($IMPERF, TYP$)
c. The boy had fished at the swamp. ($PERF, NON\_TYP$)
d. The boy was fishing at the swamp. ($IMPERF, NON\_TYP$)

[1]In such cases, the location nouns would be split by the BERT tokenizer.

Interestingly, Ferretti et al. (2007) found that atypical locations elicited significantly larger N400 amplitudes in the imperfective condition, but there were no significant differences in the perfective one, suggesting that the differences in location typicality become salient only when the event is being described as on-going. Typical locations elicited smaller N400 amplitudes, while aspect was found to have no main effect *per se*, being significant only in the interaction with typicality.

We generated another dataset, which we refer to as **Ferretti01**, by using the typicality judgements dataset created by Ferretti et al. (2001). From their data, we extracted all the verb-argument pairs for which the mean typicality rating was $>= 4$ on a Likert scale, for a total of 135 pairs, and asked two PhD students in Linguistics who were proficient English speakers, to use the same pairs to generate complete English sentences with a structure similar to the items in **Ferretti07**. A third PhD student, a native speaker of British English, made the final check of the correctness of the sentences. [2] Each item in **Ferretti01** came in the $PERF$ vs. $IMPERF$ condition (see Example 2).

(2) a. He had danced in the ballroom. ($PERF$)
b. He was dancing in the ballroom ($IMPERF$)

3.2 Model and Settings

Similarly to Misra et al. (2020), we considered the surprisal score for an argument word (in our case, the location) as a measure of the model's expectations in the given context; we replaced the location token at the end of each sentence with a *[MASK]* and we asked BERT to predict its probability.

Surprisal was shown to be an efficient predictor of self-paced reading times (Hale, 2001; Levy, 2008; Smith and Levy, 2013) and of the N400 amplitude (Frank et al., 2013), and we expected it to be inversely correlated with typicality: the more typical a location in a given context is, the less surprising it will be.

To approximate the results of the original study by Ferretti et al. (2007), BERT's surprisal scores for the locations would have to show an interaction between aspect and typicality, with the scores being significantly higher for atypical fillers only in the imperfective condition. Moreover, typical fillers should be assigned lower surprisal scores.

We experimented with the *bert-base-uncased* model, as implemented in the HuggingFace's Transformers library (Wolf et al., 2019). [3] For each sentence, we masked the location loc, corresponding to the last token in the sentence (e.g., *The girl was skating in the [MASK]*), and computed its Surprisal score $Surp$ in the context C as:

$$Surp(loc|C) = -logP(loc|C) \quad (1)$$

where $P(loc|C)$ is the probability computed by the softmax layer of BERT for the loc word as the masked token in the sentence context C.

Finally, we experimented with two different settings: a *standard* setting, in which BERT was able to see the entire context of the sentence, and a *context mask* setting, in which we used an **attention mask** on all the sentence tokens except for the verb phrase ones (see Example 3).

(3) a. The boy was fishing at the *[MASK]* (*standard*)
b. ~~The boy~~ was fishing ~~at the~~ *[MASK]* (*context mask*)

In this setting, we blocked BERT's self-attention mechanism, forcing it to use only the verb phrase to predict the masked token. In this way, we were able to analyze the effect of the tense without the interference of the other context words. [4]

4 Results and Analysis

We ran all the comparisons between scores using linear mixed effects models with the LMER function in the R statistical software (see also Appendix B for the full results). We first compared the surprisal scores that we obtained in the standard setting, in which BERT had access to all the tokens in the sentence. No main effect of aspect was found ($p > 0.1$), neither in the **Ferretti01** nor in the **Ferretti07** dataset. In other words, verb aspect did not seem to have an influence on the activation degree of typical location fillers, as the scores in the two conditions were essentially equivalent.

On the other hand, BERT's predictions for the typicality of the location fillers seemed to be extremely accurate. In the **Ferretti07** dataset, a

[2] More details on the dataset creation are in Appendix A.

[3] github.com/huggingface/transformers

[4] A similar setting was proposed by Metheniti et al. (2020) for modeling human judgements on selectional preferences: with the context mask, the authors found increased correlation between model predictions and human ratings.

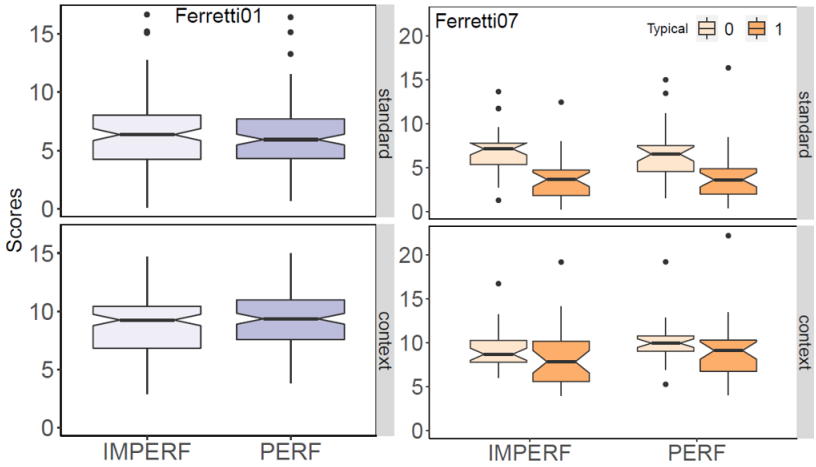

Figure 1: Boxplots with the surprisal scores calculated on the **Ferretti01** dataset (left) and on the **Ferretti07** dataset (right), in the standard setting (above) and in the context mask setting (below).

main effect of typicality was found ($p < 0.001$), with TYP sentences showing significantly lower Surprisal scores in the TYP condition than the NON_TYP ones (see also the boxplots in Figure 1, on the right). The interaction of aspect and typicality, however, was not significant ($p > 0.1$). Note that the NON_TYP locations were selected by Ferretti et al. (2007) in order to be plausible, and thus it is interesting that BERT correctly identified the ones of the TYP sentences as being more typical. However, the verb aspect did not play a role in this, since this ability was not influenced by the aspect condition. As a possible explanation, Klafka and Ettinger (2020) recently showed how the semantic information about the animacy of an argument noun in BERT was spread over the tokens of a sentence, and the same might be true also for the semantic information about the typicality of a location in a given event context. Pairwise comparisons confirmed that TYP sentences obtained significantly lower scores for both the $PERF$ and $IMPERF$ conditions ($p < 0.001$), while no significant difference between $PERF$ and $IMPERF$ in typicality condition was found. We then repeated the experiments using the context mask to block BERT's attention mechanism for all the words in the sentence except for the verb phrase tokens (recall Example 3b.) and we observed some interesting changes in our results. On the one hand, in the **Ferretti01** dataset, we observed a marginally significant effect of aspect ($p < 0.1$), with lower surprisal scores for the $IMPERF$ condition. On the other hand, in the **Ferretti07** dataset, we found main effects for *both* typicality ($p < 0.01$) and aspect ($p < 0.05$), while the interaction was again not significant ($p > 0.1$). Pairwise comparisons, similarly to the standard setting, revealed that TYP sentences had significantly lower scores for both the $PERF$ ($p < 0.05$) and the $IMPERF$ conditions ($p < 0.01$). Moreover, TYP sentences differed between $PERF$ and $IMPERF$ conditions, with the latter having significantly lower scores ($p < 0.05$). Finally, even NON_TYP sentences had lower surprisal scores in the $IMPERF$ condition, but the difference was only marginally significant ($p < 0.1$).

It should be noted that, in both settings, our results differed from those of Ferretti et al. (2007), as their study found no main effect of aspect and an interaction between aspect and typicality: atypical locations elicited significantly larger N400 components only in the imperfective condition, suggesting that imperfective verbs lead to very specific expectations on upcoming locations in human sentence processing, while expectations are way less defined with perfective verbs. By contrast, our results in the standard setting showed that BERT accurately modelled the typicality of locations without relying on the aspect of the verb, while in the context mask setting aspect influenced the predictions independently of typicality.

Finally, we checked the degree to which the typicality predictions of the model were influenced by the lexical frequencies of the target location words, which we extracted from a 2019 Wikipedia dump. [5] We found that the Surprisal scores in the **Ferretti07** dataset were not correlated at all in the standard setting, with a Spearman correlation of $\rho = -0.04$,

[5] https://github.com/IlyaSemenov/wikipedia-word-frequency.

while a weak but significant inverse correlation existed for the context mask setting at $\rho = -0.28$. This suggests that BERT's predictions are not influenced by word frequency when the entire context is available. However, when the number of contextual cues was reduced in the context mask setting, frequency might have played a more prominent role. From this point of view, it is interesting to observe the "errors" of the model: Example 4 shows the only four sentence pairs in which, in both the standard and the context mask settings, a lower score was assigned to a NON_TYP filler. Interestingly, in cases *a-c*, BERT assigns a lower Surprisal score to a more generic and frequent filler than the TYP one, which is more specific for the described event scenario. As for *d*, it can be observed that the two candidate locations (*desert-hole*) had very similar plausibility levels.

(4) a. The girl was skating/had skated in the rink (TYP) / ring (NON_TYP).
b. The boy was tobogganing/had tobogganed down the hill (TYP) / street (NON_TYP).
c. The tourist was browsing/had browsed in the shop (TYP) / park (NON_TYP).
d. The snake was slithering/had slithered in the desert (TYP) / hole (NON_TYP).

The tendency of masked language models to select a generic and frequent word when faced with the alternative of a more specific and typical filler for the event scenario was also reported by Rambelli et al. (2020) in a logical metonymy interpretation task, e.g., when asked to predict a verb for the masked position in a sentence like *The auditor begins [MASK] the taxes*, they chose generic verbs like *doing* instead of more specific ones like *auditing*, which should be preferred in the given context.

5 Conclusions

In this study, we tested whether BERT exhibited aspect-related activation effects for event locations, and whether different degrees of location typicality were identified more easily in sentences in the imperfective aspect. Verb aspect, as shown in previous studies (Ferretti et al., 2007; Madden-Lombardi et al., 2017), plays an important role in the mental representation of an event; in particular, the imperfective aspect is related to the simulation of an on-going event, giving more saliency to all the entities involved, such as the event location.

Our results showed that BERT was able to identify typical locations for events, even when it had to differentiate them from plausible but less typical ones. However, the semantic information exploited by the Transformer for the task was not linked to the verb tense, as there were no differences found between the $PERF$ and the $IMPERF$ sets. In general, no aspect-related effects on the activation of event locations were observed.

Verb aspect played a role only when the Transformer was forced to focus on the verb phrase in the context mask setting. On the one hand, BERT was still able to distinguish between typical and non-typical locations. On the other hand, the imperfective aspect was associated with significantly lower Surprisal scores for both typicality conditions. Aspect did not interact with typicality: in particular, BERT did not predict the pattern observed in Ferretti et al. (2007)'s experimental study, for which specific expectations for event locations emerge only in processing imperfective sentences.

We take these results as preliminary evidence that BERT's predictions were somewhat sensitive to aspect-related differences, and could reflect some subtle nuances of argument typicality. Our implementation is available at a public repository.[6]

Acknowledgments

We thank Lei Siyu, Chen Jing and James Britton for their help in preparing the experimental materials, and the anonymous reviewers for their feedback.

Impact Statement

Our paper is the first to study the role of verb aspect in BERT predictions on upcoming arguments. Using locations as a case study, and through a comparison with the results of the experimental studies, we showed that BERT was able to discriminate between different degrees of typicality in context. Additionally, when less cues were available from the context, aspect had an important effect on the predictions, with imperfective verbs being associated with lower surprisal values for locations. We hope that our study has shed light on new research directions to investigate the interface of grammar and event knowledge using modern Transformer-based masked language models.

[6] https://github.com/warnikchow/BERT-for-Surprisal

References

Klinton Bicknell, Jeffrey L Elman, Mary Hare, Ken McRae, and Marta Kutas. 2010. Effects of Event Knowledge in Processing Verbal Arguments. *Journal of Memory and Language*, 63(4):489–505.

Emmanuele Chersoni, Ludovica Pannitto, Enrico Santus, Alessandro Lenci, and Chu-Ren Huang. 2020. Are Word Embeddings Really a Bad Fit for the Estimation of Thematic Fit? In *Proceedings of LREC*.

Marta Coll-Florit and Silvia P Gennari. 2011. Time in Language: Event Duration in Language Comprehension. *Cognitive Psychology*, 62(1):41–79.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Allyson Ettinger. 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.

Todd R Ferretti, Marta Kutas, and Ken McRae. 2007. Verb Aspect and the Activation of Event Knowledge. *Journal of Experimental Psychology: Learning, memory, and cognition*, 33(1):182.

Todd R Ferretti, Ken McRae, and Andrea Hatherell. 2001. Integrating Verbs, Situation Schemas, and Thematic Role Concepts. *Journal of Memory and Language*, 44(4):516–547.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2013. Word Surprisal Predicts N400 Amplitude During Reading. In *Proceedings of ACL*.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural Language Models as Psycholinguistic Subjects: Representations of Syntactic State. In *Proceedings of NAACL*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of NAACL*.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of NAACL*.

Mary Hare, Michael Jones, Caroline Thomson, Sarah Kelly, and Ken McRae. 2009. Activating Event Knowledge. *Cognition*, 111(2):151–167.

Josef Klafka and Allyson Ettinger. 2020. Spying on your Neighbors: Fine-grained Probing of Contextual Embeddings for Information about Surrounding Words. In *Proceedings of ACL*.

Marta Kutas and Steven A Hillyard. 1984. Brain Potentials During Reading Reflect Word Expectancy and Semantic Association. *Nature*, 307(5947):161–163.

Roger Levy. 2008. Expectation-based Syntactic Comprehension. *Cognition*, 106(3):1126–1177.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Carol J Madden and Rolf A Zwaan. 2003. How Does Verb Aspect Constrain Event Representations? *Memory & Cognition*, 31(5):663–672.

Carol Madden-Lombardi, Peter Ford Dominey, and Jocelyne Ventre-Dominey. 2017. Grammatical Verb Aspect and Event Roles in Sentence Processing. *Plos One*, 12(12).

Yuval Marton and Asad Sayeed. 2021. Thematic Fit Bits: Annotation Quality and Quantity for Event Participant Representation. *arXiv preprint arXiv:2105.06097*.

Kazunaga Matsuki, Tracy Chow, Mary Hare, Jeffrey L Elman, Christoph Scheepers, and Ken McRae. 2011. Event-based Plausibility Immediately Influences Online Language Comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(4):913.

Ken McRae, Mary Hare, Jeffrey L Elman, and Todd Ferretti. 2005. A Basis for Generating Expectancies for Verbs from Nouns. *Memory & Cognition*, 33(7):1174–1184.

Ken McRae and Kazunaga Matsuki. 2009. People Use their Knowledge of Common Events to Understand Language, and Do So as Quickly as Possible. *Language and Linguistics Compass*, 3(6):1417–1429.

Ken McRae, Michael J Spivey-Knowlton, and Michael K Tanenhaus. 1998. Modeling the Influence of Thematic Fit (and Other Constraints) in Online Sentence Comprehension. *Journal of Memory and Language*, 38(3):283–312.

Eleni Metheniti, Tim Van de Cruys, and Nabil Hathout. 2020. How Relevant Are Selectional Preferences for Transformer-based Language Models? In *Proceedings of COLING*.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2020. Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. In *Findings of EMNLP*.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2021. Do Language Models Learn Typicality Judgments from Text? In *Proceedings of CogSci*.

Grusha Prasad, Marten Van Schijndel, and Tal Linzen. 2019. Using Priming to Uncover the Organization of Syntactic Representations in Neural Language Models. In *Proceedings of CONLL*.

Giulia Rambelli, Emmanuele Chersoni, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2020. Comparing Probabilistic, Distributional and Transformer-Based Models on Logical Metonymy Interpretation. In *Proceedings of AACL-IJCNLP*.

Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Measuring Thematic Fit with Distributional Feature Overlap. In *Proceedings of EMNLP*.

Asad Sayeed, Clayton Greenberg, and Vera Demberg. 2016. Thematic Fit Evaluation: An Aspect of Selectional Preferences. In *Proceedings of the ACL Workshop on Evaluating Vector-Space Representations for NLP*.

Nathaniel J Smith and Roger Levy. 2013. The Effect of Word Predictability on Reading Time Is Logarithmic. *Cognition*, 128(3):302–319.

TP Truitt and RA Zwaan. 1997. Verb Aspect Affects the Generation of Instrument Inferences. In *Proceedings of the Annual Meeting of the Psychonomic Society*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN Language Models Learn about Filler-Gap Dependencies? In *Proceedings of the EMNLP Workshop on Blackbox NLP*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-Art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

A Appendix A

The dataset by Ferretti et al. (2001) includes mean typicality ratings for 277 verb-location pairs on a Likert scale from 1 to 7, where 7 is the highest possible score (see examples in Table 1). The judgements were collected by asking to human subjects the following: *On a scale from 1 to 7, how common it is to **verb** in a **location**?*

Verb	Location	Mean Rating
gamble	casino	7.0
study	bedroom	5.8
marry	island	3.7
fish	pool	1.0

Table 1: Examples of the typicality judgements from the dataset by Ferretti et al. (2001). Scores range from 1 = not typical at all to 7 = very typical.

Two PhD students in Linguistics, both advanced speakers of English, voluntarily helped us in using these ratings to build the sentences of the **Ferretti01** dataset. First, we selected as typical only the verb-location pairs with a mean score $>= 4$.

Then, for each pair, we added a preposition that could be used to introduce the location in a prepositional complement. Each student took care of half of the pairs in the dataset, and then they revised each other's work. Since for this dataset we wanted BERT to compute the Surprisal scores for several candidate location fillers, the students tried to use prepositions that, given a verb, could go well with all its potential fillers in the dataset. Fillers that would have been much more likely than others given a verb-preposition pair were discarded from the dataset.

If the verb was strictly transitive, the students added a typical object, the first that came into their mind. In the end, we generated the final sentences of the dataset by randomly appending a personal pronoun subject (*He* or *She*), and we generated the sentence pairs for the two condition by varying the form of the verb: progressive past tense for the $IMPERF$ condition, past perfect tense for the $PERF$ condition.

Finally, another PhD student in Linguistics, native speaker of British English, checked that all the sentences were correct and plausible in English.

B Appendix B

	Estimates	S.E.	p
F01-Standard-Aspect	-0.24	0.28	0.39
F01-Context-Aspect	0.42	0.23	0.07 .
F07-Standard-Aspect	-0.03	0.26	0.54
F07-Standard-Typical	-3.21	0.36	< 0.001***
F07-Standard-AxT	0.38	0.52	0.46
F07-Context-Aspect	0.9	0.24	0.02*
F07-Context-Typical	-1.28	0.34	< 0.001***
F07-Context-AxT	0.24	0.48	0.62

Table 2: Results of linear mixed models on effects of aspect, typicality and the interaction of aspect and typicality (AxT), reported for both the Ferretti01 (F01) and the Ferretti07 (F07).

In Table 2, we present the tables with the full results of the linear mixed effect models for the **Ferretti01** and the **Ferretti07** dataset.

For reporting significance, we adopted the following notations: **.** for marginal significance at

$p < 0.1$, * for significance at $p < 0.05$, ** for significance at $p < 0.01$, and *** for significance at $p < 0.001$.

pairwise (F07Standard)	Estimates	S.E.	p
Impf-T:Impf-NT	-3.41	0.44	< 0.001***
Pf-T:Pf-NT	-3.02	0.44	< 0.001***
Impf-T:Pf-T	-0.16	0.36	0.97
Impf-NT:Pf-NT	0.22	0.36	0.93

Table 3: Results of pairwise comparisons in the standard setting (Impf: imperfective, Pf: perfective, T: typical, NT: non-typical).

We also report the pairwise comparisons on the **Ferretti07** dataset, both in the standard (Table 3) and in the context mask setting (Table 4).

pairwise (F07Context)	Estimates	S.E.	p
Impf-T:Impf-NT	-1.40	0.42	0.005***
Pf-T:Pf-NT	-1.16	0.42	0.029*
Impf-T:Pf-T	-1.02	0.34	0.014*
Impf:NT:Pf:NT	-0.78	0.34	0.09 .

Table 4: Results of pairwise comparisons on the context mask setting (Impf: imperfective, Pf: perfective, T: typical, NT: non-typical).