Are Frequent Phrases Directly Retrieved like Idioms? An Investigation with Self-paced Reading and Language Models

Giulia Rambelli
University of Bologna
rambelligiulia@gmail.com

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

Marco S. G. Senaldi
McGill University
marco.senaldi@mcgill.ca

Philippe Blache
Aix-Marseille University
philippe.blache@univ-amu.fr

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

Abstract

An open question in language comprehension studies is whether non-compositional multiword expressions like idioms and compositional-but-frequent word sequences are processed differently. Are the latter constructed online, or are instead directly retrieved from the lexicon, with a degree of entrenchment depending on their frequency?

In this paper, we address this question with two different methodologies. First, we set up a self-paced reading experiment comparing human reading times for idioms and both high-frequency and low-frequency compositional word sequences. Then, we ran the same experiment using the *Surprisal* metrics computed with *Neural Language Models* (NLMs).

Our results provide evidence that idiomatic and high-frequency compositional expressions are processed similarly by both humans and NLMs. Additional experiments were run to test the possible factors that could affect the NLMs' performance.

1 Introduction

It is a fact that some linguistic forms are stored in the mental lexicon, while some others have to be computed 'on the fly' by composition from smaller parts. However, the debate in linguistics and cognitive science concerns where to put the divide between 'on the fly' construction and direct retrieval (Tremblay, 2012). Theories arguing for a primary role for composition (Chomsky, 1993; Marantz, 1995; Jackendoff, 2002; Szabó, 2004) assume that rules would be responsible for the 'on the fly' computation of regular forms, while the irregular ones have to be stored in the lexicon and retrieved as a whole. On the other hand, usage-based constructionist approaches consider frequency as a crucial factor and claim that frequent forms are stored in the lexicon, while the composition mechanism is reserved to infrequent ones (Goldberg, 2003; Bybee, 2006). Accordingly, the more often a linguistic expression is encountered, the more its representation is entrenched and the easier its retrieval from the mental lexicon is (Bannard and Matthews, 2008).

The usage-based view found some strong supporting evidence in self-paced reading, EEG, and sentence recall experiments (Arnon and Snider, 2010; Tremblay and Baayen, 2010; Tremblay et al., 2011), where the speed at which highly frequent word sequences were processed suggested that they are stored and processed unitarily in the mental lexicon at least to some degree. In this research, considerable attention has been devoted to a class of recurring and conventional phrases denominated *multiword units*, *phraseological units* or *formulaic units* across different theoretical frameworks (Arnon and Snider, 2010; Siyanova-Chanturia et al., 2011; Tremblay and Baayen, 2010; Wulff, 2008; Contreras Kallens and Christiansen, 2022).

Among multiword expressions, the mechanisms underlying idiom comprehension and production have been at the core of extensive research; indeed, idioms (e.g., *break the ice*, *cut the mustard*) convey a figurative interpretation not determined by a compositional syntactic and semantic analysis of their component words (Cacciari and Tabossi, 1988; Libben and Titone, 2008; Senaldi et al., 2022). These expressions have been associated with facilitation effects in reading (Conklin and Schmitt, 2008; Titone et al., 2019) and a more positive electric signal in brain activity (Vespignani et al., 2010). To our knowledge, not many studies have directly compared the processing times of idiomatic multiword expressions and frequent

Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), pages 87–98

compositional combinations, with the exception of the study by Jolsvai et al. (2020) on three-word phrases (see Section 2.1).

In this paper, we set up a self-paced reading experiment in which we compare human reading times of English verb-determiner-noun constructions in three different conditions: **idiomatic** (*steal my thunder*), **high-frequency compositional** (*steal my wallet*) and **low-frequency compositional** (*steal my trolley*). Additionally, given the success of modern *Neural Language Models* (NLMs) and the increasing interest in using their probabilistic predictions to account for sentence processing phenomena (Futrell et al., 2018; Van Schijndel and Linzen, 2018; Wilcox et al., 2018; Michaelov and Bergen, 2020; Cho et al., 2021; Michaelov and Bergen, 2022a; Michaelov et al., 2023), we repeated the experiment by extracting the *Surprisal* values (Hale, 2001; Levy, 2008) of the words in the stimuli with several RNN- and Transformer-based models, to compare them with the human results. We chose this measure because Surprisal is considered an indicator of the processing load associated with a word; experiments have found a strong correlation between biometric and computational values (Ryu and Lewis, 2021).

Our results show that **humans process idiomatic and high-frequency compositional expressions significantly faster than low-frequency compositional ones** and, in parallel, NLMs assign to them significantly lower Surprisal values. Among the models we tested, we found out that **the smaller version of GPT2 and a 2-layer LSTM obtained the exact same score patterns as human subjects**; we observed no significant difference between the Surprisal scores in the idiomatic and the high-frequency conditions, but the values for the infrequent condition were significantly higher.[1]

2 Related Work

2.1 Direct access of Idiomatic and Frequent Sequences

The idea that frequently-occurring multiword expressions may be stored and processed holistically had been put forth already by Biber et al. (2000). Tremblay et al. (2011) set up a self-paced reading experiment comparing frequent lexical bundles (e.g., *whatever you think about it*) and lower-frequency control sequences (e.g., *whatever you do about it*), and they found that the former were read faster by human subjects across different experimental settings. Arnon and Snider (2010) compared the reaction times in phrasal decision tasks between frequent and infrequent word sequences (e.g., *I don't know why* vs. *I don't know who*), where the subparts of the sequence were matched for frequency, and they reported a clear effect of phrase frequency on recognition times. Tremblay et al. (2011) described a four-word production task in which the participants had to say the word sequences that were shown to them, and their production onset latencies and total durations were measured. The authors found several main effects related to word frequencies, contextual predictability, and mutual information, deemed as indicative of the holistic storage of forms.

Among multiword expressions, it is generally acknowledged that idiomatic constructions play a special role, as they convey a figurative meaning that cannot be accessed by merely combining the semantics of their components (*non-compositionality*; (Jackendoff, 2002)). Converging evidence from online methodologies supports facilitation in processing for idioms with respect to non-idiomatic phrases (Cacciari and Tabossi, 1988; Conklin and Schmitt, 2008; Vespignani et al., 2010; Siyanova-Chanturia et al., 2011; Titone et al., 2019). There is an open debate about how idioms are represented in the mental lexicon and processed during comprehension: while the *non-compositional* view considers idioms as frozen strings directly accessed during comprehension (Swinney and Cutler, 1979; Cacciari and Tabossi, 1988, i.a.), recent evidence suggests that idiom comprehension involves both direct meaning retrieval and compositional analysis at different comprehension stages, thus validating hybrid models of idiom processing (Libben and Titone, 2008; Titone et al., 2019).

In particular, hybrid views predict that an idiom's degree of familiarity or subjective frequency modulates the availability of direct retrieval as a processing strategy. Indeed, prior studies had shown speakers to engage in a more compositional processing strategy when idioms are less frequent or familiar, for example, because they appear in a non-canonical modified form or they are being processed in a second language (Senaldi and Titone, 2022; Senaldi et al., 2022). Vice versa, a question that remains unaddressed is whether frequent but compositional word combinations can benefit

[1] Data and code available at: `https://osf.io/4jg2b/?view_only=e3679a4df4c248fb8819156b392e92ad`.

from some form of direct memory access during processing.

Jolsvai et al. (2020), to our knowledge, is the only study attempting a comparison between three-word idiomatic expressions, frequent compositional phrases, and fragments. A phrasal decision task revealed that the meaningfulness of the chunk sped up reaction times, which were similar for idioms (*play the field*) and frequent phrases (*nothing to wear*), while phrasal fragments (*without the primary*) took considerably more time. However, the stimuli across the three conditions were just matched on sub-components' frequency, without any constraint about the superficial realization of the constructions. Unlike Jolsvai and colleagues, we only focused on English verb constructions. We manipulated frequency and degree of compositionality by changing the direct object while keeping the verb constant. Across experimental conditions, the same verb could appear in an idiom (*spill the beans*), a high-frequency compositional construction (*spill the milk*), and a low-frequency compositional construction (*spill the rice*, see Section 3.1).

2.2 Constructions and Idioms in Transformer Language Models

With the rise to the popularity of Transformer language models in NLP (Vaswani et al., 2017; Devlin et al., 2019), several studies explored the nature of the linguistic representations of Transformers and how they handle compounds and other types of non-compositional expressions (Shwartz and Dagan, 2019; Rambelli et al., 2020; Garcia et al., 2021a,b; Dankers et al., 2022). Interestingly, some studies specifically used the probing paradigm to analyze to what extent Transformers have access to construction knowledge (Weissweiler et al., 2023; Pannitto and Herbelot, 2023), and there is a general agreement that they have some knowledge about the formal/syntactic aspects of constructions (Madabushi et al., 2020; Weissweiler et al., 2022). In contrast, the evidence about the encoding of meaning aspects is mixed, depending on the specific constructions and the type of semantic knowledge being probed (Li et al., 2022; Weissweiler et al., 2022). This literature primarily focused on analyzing idioms and constructions at the level of the Transformer representations.

To our knowledge, there have been no attempts yet to model the effects of such linguistic expressions on human sentence processing, for example, in terms of reading times or eye-tracking fixations. In computational psycholinguistics, it has become common to use NLMs to extract word Surprisals (Hale, 2001; Levy, 2008) and use such values to model human behavioral patterns. For instance, Transformer Surprisal has been shown to accurately predict human reading times from naturalistic reading experiments, outperforming the metrics derived from architectures based on recurrent neural networks (Wilcox et al., 2020; Merkx and Frank, 2021). Evaluating computational models on sentence processing data is, in our view, a necessary complement to the construction probing tasks, as it makes it possible to test the predictions against the cognitively-plausible benchmark represented by human behavior (Rambelli et al., 2019).

3 Experiment 1: Self-paced Reading (SPR)

3.1 Stimuli and SPR Data

Stimuli consisted of 48 English verb-determiner-noun phrases appearing in 3 experimental conditions, namely as idiomatic expressions (ID, *spill the beans*), high-frequency compositional phrases (HF, *spill the milk*) and low-frequency compositional phrases (LF, *spill the rice*). The three conditions shared the same verb. First, we selected all verb-determiner-noun expressions from two normative datasets of American English idioms (Libben and Titone, 2008; Bulkes and Tanner, 2017) and Kyriacou et al. (2020)'s study. To generate matched HF and LF compositional phrases for each of the items, we relied on the enTenTen18 corpus (Jakubíček et al., 2013), a large part-of-speech parsed corpus of English made up of texts collected from the Internet (21.9 billion words). We employed the sketchEngine[2] tools (Kilgarriff et al., 2014) to run our queries. We verified that the HF expression had a comparable log frequency with the corresponding idiom and that the noun-verb association score was similar or larger than the association score in the idiomatic phrase, relying on the LogDice score implemented in SketchEngine (Rychlý, 2008). Moreover, we matched the nouns in all three conditions for log-transformed frequency and character word length. We discarded the idioms for whom finding an appropriate matched HF was impossible. Finally, we ran an *Idiom Familiarity* survey to exclude unfamiliar idioms, and a *Typical Objects Production* study, to verify that the noun in the low-

[2] http://www.sketchengine.eu

Cond.	Context	Precritical region - **Critical region** - Postcritical region
ID	Finn changed his life after his father's death.	All of a sudden he **kicked the habit** and stopped smoking cigarettes.
HF	It was the first match for Finn.	All of a sudden he **kicked the ball** into the net and won the match.
LF	That day, Finn had completely lost his temper.	All of a sudden he **kicked the sister** of his best friend in the head.

Table 1: Example of the stimuli for the self-paced reading experiment.

frequent Condition was not in the list. We collected online judgments from 57 and 74 North American subjects, respectively. Idioms receiving a familiarity score lower than 4 were left out. The final selection led to 48 triplets consisting of a highly familiar idiom and matched frequent and infrequent compositional bigrams.

From the bigram list, we built the experimental stimuli. Specifically, a stimulus consisted of a sentence containing a contextual preamble displayed as a whole and a sentence containing the target phrase[3] presented word-by-word using the moving-window SPR paradigm (see Table 1). Stimuli were split into three counterbalanced lists such that only one condition of the triple was in a list[4], and they were randomized for each participant. The experiment was delivered remotely, and participants were recruited using Prolific [2021].[5] We collected responses from 90 subjects from the United States and Canada, all self-reported L1 speakers of English aged between 18 and 50. We considered the reading times (henceforth RTs) on the object noun, that is, the last word of the target bigram. We removed responses of less than 100 ms (Jegerski, 2013) as well as reading times that were 2.5 standard deviations above each condition's mean, resulting in 7.3% data loss. Then, we ran a linear mixed model in R (v. 3.6.3) with the `lme4` package (Bates et al., 2015). We included log-transformed RTs as the dependent variable, while the condition, the noun length, the verb frequency (log-transformed), and the trial number were entered in the models as fixed effects. Finally, the Subject and Item were treated as random effects. Significance was computed using the *lmerTest* package (Kuznetsova et al., 2017), which applies Satterthwaite's method to estimate degrees of freedom and generates p-values for mixed models.

[3]Context and target sentences were manually created by the authors and validated by an English teacher.

[4]It is a common methodology in psycholinguistics to prevent possible priming effects.

[5]www.prolific.co

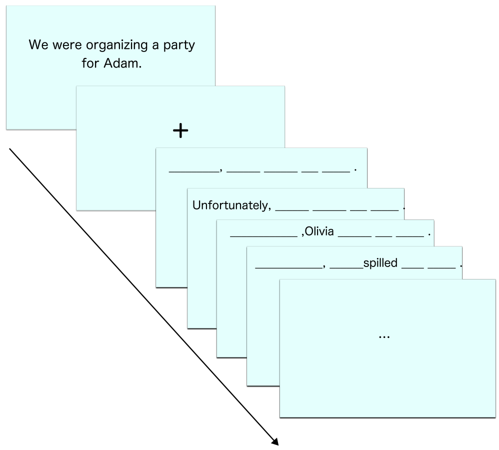

Figure 1: SPR procedure. 1) A context sentence appears in the center of the screen; the participant goes to the next sentence by pressing the space key. 2) The target text is displayed as a series of dashes on the screen, each dash representing a character. The first word appears when the participant presses the space key, replacing the corresponding dashes. Each button presses cause the previous words to be overridden again by dashes during the current word surface.

3.2 Results

The difference in RTs between ID and HF turned out to be not statistically significant (β= .002594, $t = .191$, $p = .85$), while it was significantly different between ID and LF (β = .02982, $t = 2.190$, $p = .0299*$). When mapping the HF condition to the intercept, there was still a statistically significant difference between HF and LF ($\beta = .0272$, $t = 2.007$, $p = .0466*$). To be consistent with common practices in the psycholinguistic literature, we included the trial number as a fixed effect: as expected, RTs at the end of the experiment tended to be shorter than at the beginning.

Analyses revealed no significant differences in reading times between idioms with a non-compositional meaning and high-frequency compositional phrases; there was facilitation in both conditions, compared to low-frequency compositional phrases. Although reading times do not allow to draw conclusions on how these phrases are rep-

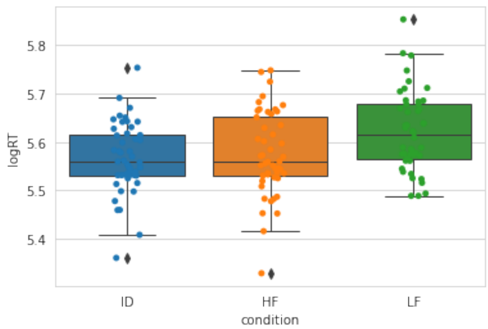

Figure 2: RTs distribution across the conditions.

resented at the brain level, the collected evidence seems in line with the claims of usage-based constructionist models (Goldberg, 2006; Wulff, 2008; Bybee, 2010). Accordingly, frequency of exposure determines the degree of lexical entrenchment of non-compositional and compositional structures alike; thus, even highly frequent compositional structures can end up being represented as wholes in the lexicon without being necessarily composed piecemeal during online processing.

Since our results reveal comparable processing times between HF and ID phrases and there is consistent evidence that idioms are at least to some extent retrieved directly from memory during processing, we can hypothesize a similar processing strategy to be at play for both. Another explanation is that since ID and HF phrases are frequently encountered by speakers, they are read faster because the processing system relies on analogical similarities with a high number of stored exemplars (Ambridge, 2020; Rambelli et al., 2022). Finally, RTs for infrequent phrases were significantly slower, even if the edge on ID and HF was relatively small: we presume that the information introduced in context sentences plays a role in reducing the effort to interpret less predictable expressions.

4 Experiment 2: Modeling Reading Times with Neural Language Models (NLMs)

4.1 NLM Architectures

To investigate which NLM architecture explains SPR data, we chose Transformers and recurrent networks (RNN), which are traditionally ascribed as a cognitively plausible model of human sentence processing (Elman, 1990). RNNs are inherently sequential: a token's representation depends on the previous hidden state to form a new hidden state. In contrast, Transformers have a self-attention layer allowing to 'attend' to parts of previous input directly (Vaswani et al., 2017).

Among the Transformers, we tested both autoregressive models (i.e., GPT), where the probability of the target word is computed based on the left context, and bidirectional models (like BERT (Devlin et al., 2019)) that instead predict a word looking at both the left and right context. **GPT2** (Radford et al., 2019) is a unidirectional Transformer LM pre-trained on WebText for a total of 8 million documents of data (40 GB) and has a vocabulary size of 50.257. We employed all four versions of GPT-2 (small/medium/large/xl) for our experiments to test if the model size has an impact on the results (parameters are reported in Appendix A). Unlike GPT2, **BERT** (Devlin et al., 2019) was the first to adopt the bidirectional training of Transformer for a language modeling task. It is trained both on a masked language modeling task (i.e., the model attempts to predict a masked token based on the surrounding context) and on a next sentence prediction task, as the model receives sentence pairs in input and has to predict whether the second sentence is subsequent to the first one in the training data. BERT has been trained on a concatenation of the BookCorpus and the English Wikipedia for a total of around3300M tokens. We used the `bert-base-uncased` pre-trained version in our experiments. In addition, we selected the Text-To-Text Transfer Transformer (**T5**) (Raffel et al., 2020), an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format. We experimented with the T5-base model (220 million parameters), trained on a 7 TB dataset. All models were loaded through `minicons` (Misra, 2022),[6] a Python library facilitating the probability computations with the LMs that are accessible through the `transformers` package by HuggingFace.

Moreover, we compared Transformers with two kinds of recurrent networks as a baseline. **TinyLSTM** is a two-layer LSTM recurrent neural network trained with a next-word prediction on the Wikitext-2 dataset, a collection of over 100 million tokens (Stephen et al., 2017). **GRNN** is the best-performing model described in the supplementary materials of Gulordava et al. (2018). It was trained on 90 million tokens of English Wikipedia with two hidden layers of 650 hidden units. Both models

[6] https://github.com/kanishkamisra/minicons

	ID_{median}	HF_{median}	LF_{median}	ID-HF	ID-LF	HF-LF
GPT2-small	5.36 (IQR 4.82)	6.43 (IQR 3.57)	12.7 (IQR 5.19)	ns	***	***
GPT2-medium	4.59 (IQR 4.60)	6.66 (IQR 5.58)	12.2 (IQR 4.61)	*	***	***
GPT2-large	3.96 (IQR 4.90)	6.71 (IQR 5.93)	12.4 (IQR 4.64)	*	***	***
GPT2-xl	2.41 (IQR 3.98)	4.46 (IQR 3.00)	8.00 (IQR 4.05)	*	***	***
BERT-base-uncased	21.6 (IQR 4.68)	20.1 (IQR 7.12)	21.5 (IQR 4.7)	ns	ns	ns
T5-base	18.5 (IQR 4.32)	17.1 (IQR 5.17)	20.1 (IQR 6.5)	ns	ns	**
TinyLSTM	11.8 (IQR 2.98)	11.7 (IQR 5.28)	14.1 (IQR 3.69)	ns	***	**
GRNN	12.0 (IQR 5.23)	9.60 (IQR 4.02)	14.2 (IQR 4.74)	*	***	***

Table 2: Comparison of Surprisal scores using Wilcoxon Signed-Rank Test (with Bonferroni's correction). *p= < .05, **p= < .01, and ***=p < .001.

were queried with the Language Model Zoo,[7] an open-source repository of state-of-the-art language models, designed to support black-box access to model predictions (Gauthier et al., 2020).

4.2 Methodology

Reading times are a common way to identify readers' facilitation effects in comprehension. For NLMs, we measured the Surprisal of the next word, which is notoriously an important predictor of reading times in humans (Smith and Levy, 2013) and has been largely used to test language models' abilities (cf. Section 2.2).

The Surprisal of a word *w* (Hale, 2001; Levy, 2008) is defined as the negative log probability of the word conditioned on the sentence context

$$Surprisal(w) = -logP(w|context) \quad (1)$$

where the context can be words on the left (for autoregressive models) or words both on the left and on the right of the target *w*. We passed the stimuli sentences presented in the previous experiment to all selected NLMs and computed the Surprisal of the object noun in each experimental condition. The Surprisal score should reveal how easy it is to process a target word: the lower the score, the higher the facilitation effect. For out-of-vocabulary words, we computed the sum of the Surprisals of the subtokens.

4.3 Results of Surprisal Analyses

Table 2 summarizes the difference among conditions for each model. We compared the Surprisal distribution in the three conditions by relying on the non-parametric Wilcoxon signed rank test with the Bonferroni correction. We applied the `wilcoxon_test` function from the `rstatix` package in R language. The Wilcoxon test shows a statistical difference between the Surprisals of ID and HF conditions ($p < .05$), differently than in human reading times. Specifically, all the GPT2 models, with the exception of the 'small' version, produce lower scores for ID condition than for HF. This outcome seems to indicate that the idiomatic expression is more expected by the model, even if we controlled the stimuli to have a similar bigram frequency and verb-noun association. Surprisingly, the other Transformer model shows an opposite trend: BERT-base-uncased and T5-base have an average Surprisal of HF lower than those for ID condition, and there is no significant difference not only between ID and HF conditions but also between ID and LF. This outcome, confirmed by the boxplot visualization (Figure 3), reveals that bidirectional models are not sensitive to the difference among the three conditions. Moreover, the scores are consistently higher than GPT2 models, indicating that all the expressions are quite unexpected by the two Transformer architectures.

Considering recurrent networks, GRNN performs similarly to the (larger) T5-base model: the average Surprisal of HF is lower than those for ID condition. However, in this case, HF scores are significantly lower than ID. We could infer that this recurrent neural network prefers the frequent compositional competition, while it is more surprised by the same frequent but figurative expression.

There are only two models whose Surprisals are comparable to human RTs: **GPT2-small** and **tinyLSTM**. The fact that the smaller GPT2 model resembles human performance is interesting and might be further evidence of the *inverse scaling* that has been observed in LMs for several natural language phenomena; that is, the more the model size grows, the less human-like its behavior is (Wei et al., 2022; Michaelov and Bergen, 2022b; Oh and Schuler, 2022; Jang et al., 2023). Oh and Schuler (2022) suggested that this behavior can be explained by the fact that larger LMs have seen many more word sequences than humans; as model size grows, the

[7] `https://github.com/cpllab/lm-zoo`

predictions tend to be more and more accurate for open class words, to the point of underestimating their reading time delays.

We found no statistical correlation between the human RTs with the NLMs' Surprisals, as it is evident from the scatterplots in Figure 5 (analyses were conducted using the Spearman's correlation).

4.4 The Role of Context

The results of the SPR experiment revealed that, while there is a significant difference between ID/HF conditions and infrequent phrases, the advantage is relatively small (in milliseconds). A plausible explanation is that the preceding context has a priming effect on the noun interpretation in the target sentence, regardless of the condition. As an additional investigation, we re-run all models but fed them only with the target sentence without the contextual sentence. A two-way ANOVA was performed to analyze the effect of Condition and Context on Surprisal scores for all models. For a visual comparison, we plotted the Surprisal distribution obtained both with and without the context sentence (Figure 3). This analysis reveals that **recurrent neural networks** (tinyLSTM and GRNN) **and bidirectional models** (BERT and T5) **produce the same Surprisal with or without the context sentence**. Two-way ANOVA revealed no statistically significant interaction between the effects of Condition and Context (BERT: $F = .001$, $p = .97$; T5: $F = .016$, $p = .899$; tinyLSTM: $F = .343$, $p = .559$; GRNN: $F = .014$, $p = .905$). Simple main effects analysis showed that Context did not have a statistically significant effect, while Condition did have a statistically significant effect on Surprisal scores ($p < .001$). This outcome suggests that, for all these models, word prediction is highly localized, and the preceding context has little or no priming effect on the expectation of the next word. This evidence could also explain BERT and T5-base performances: a word's expectancy is not affected by the preceding context, thus the model is highly surprised by all words, regardless of verb associations (frequent or infrequent bigram) and expression type (idiomatic or literal). However, this observation should be further verified with more targeted experiments.

Contrarily, we observe the expected trend for all **GPT2 models**: **Surprisal scores decrease, giving a context sentence before the stimuli**. The two-way ANOVA revealed that there was not a

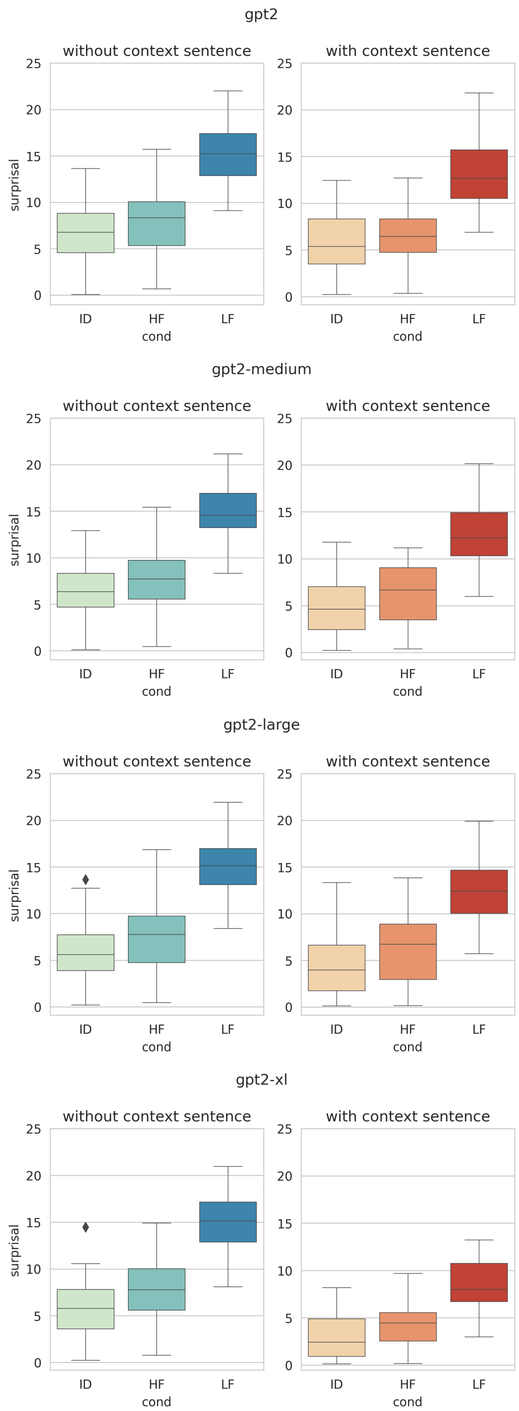

Figure 3: Surprisal distributions per conditions for GPT2 models, with (right) and without (left) the context sentence. The comparison of boxplots reveals that Surprisal scores decrease by giving a context sentence before the stimuli.

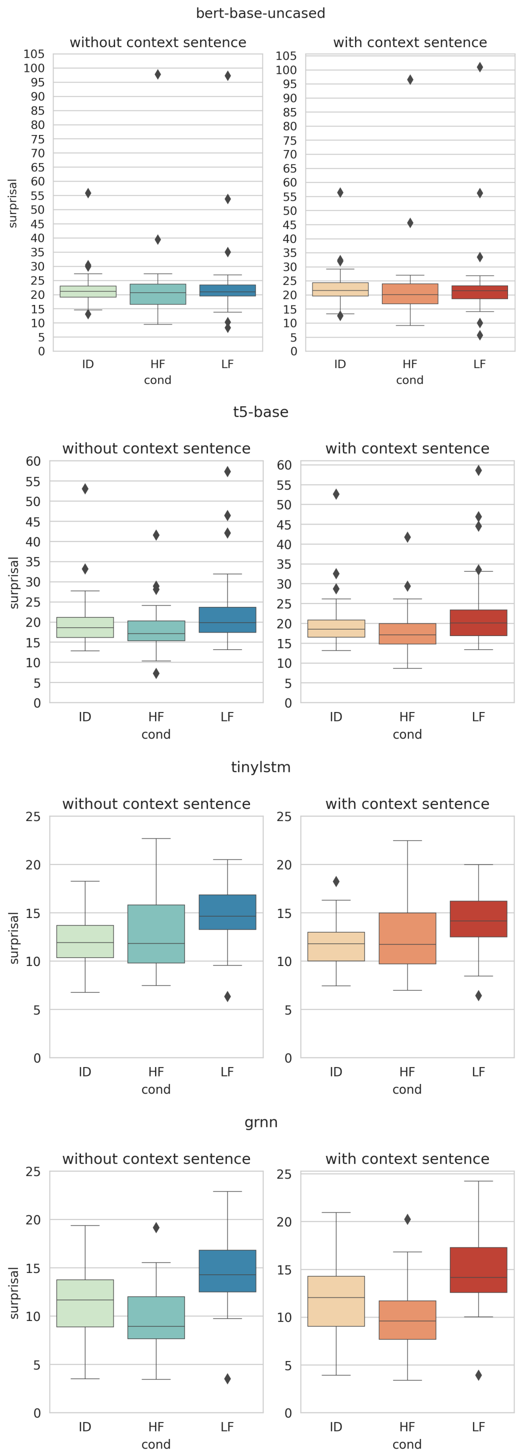

Figure 4: Surprisal distributions per conditions for BERT-base-uncased, T5-base, tinyLSTM, and GRNN, with and without the context sentence. The comparison of boxplots reveals that Surprisal scores are the same regardless the context.

statistically significant interaction between Context and Condition for all variants, with the exception of GPT2-xl (GPT2: $F = .014$, $p = .905$; GPT2-medium: $F = .883$, $p = .348$; GPT2-large: $F = 1.351$, $p = .246$; GPT2-xl: $F = 106.49$, $p < .001***$). However, Context as a simple main effect does have a statistically significant effect in all models (GPT2: $F = 9.559$, $p = .002**$; GPT2-medium: $F = 14.686$, $p < .001***$; GPT2-large: $F = 15.398$, $p < .001***$; GPT2-xl: $F = 8.31$, $p = .004**$). What is important to notice, however, is that the differences among the conditions are kept constant. Accordingly, GPT2 models show LF condition is less expected than the other two, and Surprisal values for idioms and high-frequent expressions are similar independently of the context. This outcome is important because it tells us that, even if the context has a facilitatory effect on LMs' processing, it is not the main cause for Surprisal scores.

5 Discussion

This study is part of a broad research about how people access meaning during language processing and to what extent NLMs replicate human behavior. In our view, comparing idioms to frequent literal expressions may provide novel insights into the influence of phrase frequency on language processing and the integration of compositional and noncompositional mechanisms.

In the SPR experiment, we found that people read idioms and frequent compositional units at comparable speeds. The results of this study require further investigation. For instance, we could analyze the influence of context on comprehension by collecting reading times of the stimuli presented without the contextual sentence; as well, we could present the same stimuli in an eye-tracking paradigm to record more fine-grained measures than mere reading time. Secondly, instead of relying only on corpus frequencies, we could explore the relationship between reading times and other ratings, such as cloze probability, plausibility, or meaningfulness (Jolsvai et al., 2020). Moreover, we restricted this study to N-det-V pattern, but we are planning to apply the experiment to other types of multiword expressions. Finally, we are planning to extend this investigation to other languages to assess the cross-linguistic validity of our findings.

The experimental evidence provided by the computational experiment confirms our behavioral find-

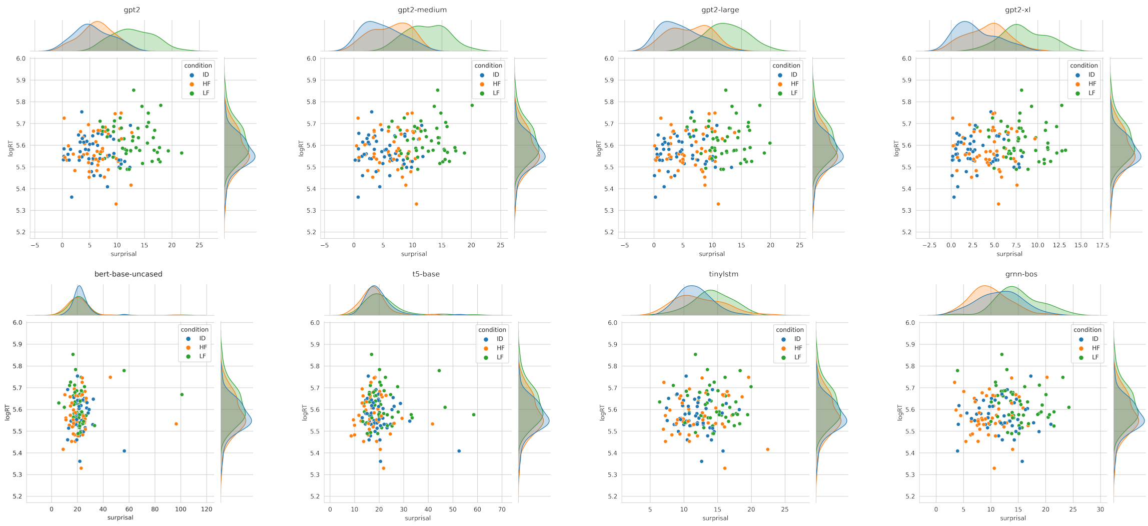

Figure 5: Scatterplot showing the relationship between Surprisal scores (x-axis) and RTs (y-axis).

ings: both idiomatic and frequent expressions are highly expected by GPT2 models. Interestingly, the models that mirrored more closely human reading patterns are the smallest ones, in agreement with the findings recently reported by the literature on *inverse scaling* in NLMs. Future research includes replicating this study with other architectures, including the successor of GPT2, namely GPT3.

A compelling behavior of NLMs regards the role of context: it seems to affect little or not at all the Surprisal scores. This evidence suggests that the Surprisal of a word depends more on the ease of access to a word in the vocabulary than on the semantic integration with previous words. In other words, frequent expressions might be 'memorized' and easily retrieved, and context words do not show relevant priming effects. We plan to investigate this outcome in future experiments and verify how humans react without the contextual sentence. Besides, we can conclude that the converging evidence from humans and LMs suggests that multiword expressions, both idiomatic and compositional ones, are processed more holistically than compositionally.

Our experiment opens up to many possibilities for further analyses and refinements. For example, considering the behavioral experiment, a peculiarity of our design is that the point at which an idiom becomes recognizable is located at the end of the target phrase. Even if reading times on this specific word gives us insight into the facilitation access to construction meaning, the cognitive effort in processing that word is not limited to the word itself but could emerge in the subsequent text (*spillover* effect; Rayner and Duffy (1986); Reichle et al. (2003)). Considering the computational experiment, we just analyzed the probability output of a target word through the Surprisal scores, but in the future, it would be useful to adopt interpretability techniques to get more insights on the hidden representations of the NLMs (Yin and Neubig, 2022; Belrose et al., 2023).

We hope that our findings can contribute to the existing research in multiword expression processing, paving the way for forthcoming studies on how the compositional and noncompositional mechanisms alternate during interpretation.

Limitations

An obvious limitation is that our analysis was limited to English, and we hope to replicate the same experimental design for other languages in the future. Moreover, we limited ourselves to just one type of construction (verb phrases).

Acknowledgements

A special acknowledgment goes to Christelle Zielinski for her support in implementing the online behavioral experiment and data explorations. We would also like to thank the reviewers for their insightful feedback.

This research was supported by grants ANR-16-CONV-0002 (ILCB) and the Excellence Initiative of Aix-Marseille University (A*MIDEX), and by the General Research Fund (B-Q0AH) at the Hong Kong Polytechnic University.

References

Ben Ambridge. 2020. Against Stored Abstractions: A Radical Exemplar Model of Language Acquisition. *First Language*, 40(5-6):509–559.

Inbal Arnon and Neal Snider. 2010. More than Words: Frequency Effects for Multi-word Phrases. *Journal of Memory and Language*, 62(1):67–82.

Colin Bannard and Danielle Matthews. 2008. Stored Word Sequences in Language Learning: The Effect of Familiarity on Children's Repetition of Four-word Combinations. *Psychological Science*, 19(3).

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112*.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 2000. *Longman Grammar of Spoken and Written English*. Longman London.

Nyssa Z. Bulkes and Darren Tanner. 2017. "Going to town": Large-scale Norming and Statistical Analysis of 870 American English Idioms. *Behavior Research Methods*, 49(2):772–783.

Joan Bybee. 2006. From Usage to Grammar: The Mind's Response to Repetition. *Language*, pages 711–733.

Joan L. Bybee. 2010. *Language, Usage and Cognition*. Cambridge University Press.

Cristina Cacciari and Patrizia Tabossi. 1988. The Comprehension of Idioms. *Journal of Memory and Language*, 27(6):668–683.

Won Ik Cho, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Modeling the Influence of Verb Aspect on the Activation of Typical Event Locations with BERT. In *Findings of ACL-IJCNLP*.

Noam Chomsky. 1993. A Minimalist Program for Linguistic Theory. *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*.

Kathy Conklin and Norbert Schmitt. 2008. Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers? *Applied Linguistics*, 29(1):72–89.

Pablo Contreras Kallens and Morten H Christiansen. 2022. Models of Language and Multiword Expressions. *Frontiers in Artificial Intelligence*, 5:24.

Verna Dankers, Christopher G Lucas, and Ivan Titov. 2022. Can Transformer be too Compositional? Analysing Idiom Processing in Neural Machine Translation. In *Proceedings of ACL*.

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*.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. RNNs as Psycholinguistic Subjects: Syntactic State and Grammatical Dependency. *arXiv preprint arXiv:1809.01329*.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the Representations of Idiomaticity in Vector Models with a Noun Compound Dataset Labeled at Type and Token Levels. In *Proceedings of ACL*.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for Idiomaticity in Vector Space Models. In *Proceedings of EACL*.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In *Proceedings of ACL: Demo*.

Adele E Goldberg. 2003. Constructions: A New Theoretical Approach to Language. *Trends in Cognitive Sciences*, 7(5):219–224.

Adele E. Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press on Demand.

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*.

Ray Jackendoff. 2002. *Foundations of Language*. Oxford University Press.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The tenten corpus family. In *7th International Corpus Linguistics Conference CL*, pages 125–127. Lancaster University.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can Large Language Models Truly Understand Prompts? A Case Study with Negated Prompts. In *Transfer Learning for Natural Language Processing Workshop*, pages 52–62. PMLR.

Jill Jegerski. 2013. Self-paced Reading. In *Research Methods in Second Language Psycholinguistics*, pages 36–65. Routledge.

Hajnal Jolsvai, Stewart M McCauley, and Morten H Christiansen. 2020. Meaningfulness Beats Frequency in Multiword Chunk Processing. *Cognitive Science*, 44(10):e12885.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten Years On. *Lexicography*, 1(1):7–36.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune HB Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(1):1–26.

Marianna Kyriacou, Kathy Conklin, and Dominic Thompson. 2020. Passivizability of Idioms: Has the Wrong Tree Been Barked Up? *Language and speech*, 63(2):404–435.

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

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural Reality of Argument Structure Constructions. In *Proceedngs of ACL*.

Maya R. Libben and Debra A. Titone. 2008. The Multi-determined Nature of Idiom Processing. *Memory & Cognition*, 36(6):1103–1121.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT Meets Construction Grammar. In *Proceedings of COLING*.

Alec Marantz. 1995. The Minimalist Program. In *The Principles and Parameters Approach to Linguistic Theory*, pages 351–382. Blackwell.

Danny Merkx and Stefan L Frank. 2021. Human Sentence Processing: Recurrence or Attention? In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

James A Michaelov and Benjamin K Bergen. 2020. How Well Does Surprisal Explain N400 Amplitude under Different Experimental Conditions? In *Proceedings of CONLL*.

James A Michaelov and Benjamin K Bergen. 2022a. Collateral Facilitation in Humans and Language Models. In *Proceedings of CONLL*.

James A Michaelov and Benjamin K Bergen. 2022b. 'Rarely' a Problem? Language Models Exhibit Inverse Scaling in their Predictions Following 'Few'-type Quantifiers. *arXiv preprint arXiv:2212.08700*.

James A Michaelov, Seana Coulson, and Benjamin K Bergen. 2023. Can Peanuts Fall in Love with Distributional Semantics? *arXiv preprint arXiv:2301.08731*.

Kanishka Misra. 2022. minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint arXiv:2203.13112*.

Byung-Doh Oh and William Schuler. 2022. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? In *Proceedings of EMNLP*.

Ludovica Pannitto and Aurélie Herbelot. 2023. CALaMo: A Constructionist Assessment of Language Models. *arXiv preprint arXiv:2302.03589*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer. *Journal of Machine Learning Research*, 21(1):5485–5551.

Giulia Rambelli, Emmanuele Chersoni, Philippe Blache, Chu-Ren Huang, and Alessandro Lenci. 2019. Distributional Semantics Meets Construction Grammar. Towards a Unified Usage-based Model of Grammar and Meaning. In *Proceedings of the ACL Workshop on Designing Meaning Representations*.

Giulia Rambelli, Emmanuele Chersoni, Philippe Blache, and Alessandro Lenci. 2022. Compositionality as an Analogical Process: Introducing ANNE. In *Proceedings of the AACL-IJCNLP Workshop on Cognitive Aspects of the Lexicon*.

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*.

Keith Rayner and Susan A Duffy. 1986. Lexical Complexity and Fixation Times in Reading: Effects of Word Frequency, Verb Complexity, and Lexical Ambiguity. *Memory & Cognition*, 14(3):191–201.

Erik D. Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The EZ Reader Model of Eye-Movement Control in Reading: Comparisons to Other Models. *Behavioral and Brain Sciences*, 26(4):445–476.

Pavel Rychlý. 2008. A Lexicographer-Friendly Association Score. In *RASLAN*, pages 6–9.

Soo Hyun Ryu and Richard L Lewis. 2021. Accounting for Agreement Phenomena in Sentence Comprehension with Transformer Language Models: Effects of Similarity-based Interference on Surprisal and Attention. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Marco S. G. Senaldi, Junyan Wei, Jason W Gullifer, and Debra Titone. 2022. Scratching your Tête over Language-switched Idioms: Evidence from Eye-movement Measures of Reading. *Memory & Cognition*, 50(6):1230–1256.

Marco S.G. Senaldi and Debra Titone. 2022. Less Direct, More Analytical: Eye-Movement Measures of L2 Idiom Reading. *Languages*, 7(2):91.

Vered Shwartz and Ido Dagan. 2019. Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Anna Siyanova-Chanturia, Kathy Conklin, and Norbert Schmitt. 2011. Adding More Fuel to the Fire: An Eye-tracking Study of Idiom Processing by Native and Non-native Speakers. *Second Language Research*, 27(2):251–272.

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

Merity Stephen, Xiong Caiming, Bradbury James, and Richard Socher. 2017. Pointer Sentinel Mixture Models. *Proceedings of ICLR*.

David A Swinney and Anne Cutler. 1979. The Access and Processing of Idiomatic Expressions. *Journal of Verbal Learning and Verbal Behavior*, 18(5):523–534.

Zoltán Gendler Szabó. 2004. Compositionality. *Stanford Encyclopedia of Philosophy*.

Debra Titone, Kyle Lovseth, Kristina Kasparian, and Mehrgol Tiv. 2019. Are Figurative Interpretations of Idioms Directly Retrieved, Compositionally Built, or Both? Evidence from Eye Movement Measures of Reading. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 73(4):216.

Antoine Tremblay. 2012. *Empirical Evidence for an Inflationist Lexicon*. De Gruyter.

Antoine Tremblay and R Harald Baayen. 2010. Holistic Processing of Regular Four-word Sequences: A Behavioral and ERP Study of the Effects of Structure, Frequency, and Probability on Immediate Free Recall. *Perspectives on Formulaic Language: Acquisition and Communication*, pages 151–173.

Antoine Tremblay, Bruce Derwing, Gary Libben, and Chris Westbury. 2011. Processing Advantages of Lexical Bundles: Evidence from Self-paced Reading and Sentence Recall Tasks. *Language Learning*, 61(2):569–613.

Marten Van Schijndel and Tal Linzen. 2018. Modeling Garden Path Effects without Explicit Hierarchical Syntax. In *Proceedings of CogSci*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.

Francesco Vespignani, Paolo Canal, Nicola Molinaro, Sergio Fonda, and Cristina Cacciari. 2010. Predictive Mechanisms in Idiom Comprehension. *Journal of Cognitive Neuroscience*, 22(8):1682–1700.

Jason Wei, Yi Tay, and Quoc V Le. 2022. Inverse Scaling Can Become U-shaped. *arXiv preprint arXiv:2211.02011*.

Leonie Weissweiler, Taiqi He, Naoki Otani, David R Mortensen, Lori Levin, and Hinrich Schütze. 2023. Construction Grammar Provides Unique Insight into Neural Language Models. *arXiv preprint arXiv:2302.02178*.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The Better Your Syntax, the Better Your Semantics? Probing Pretrained Language Models for the English Comparative Correlative. In *Proceedings of EMNLP*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What Do RNN Language Models Learn about Filler-gap Dependencies? *arXiv preprint arXiv:1809.00042*.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. *arXiv preprint arXiv:2006.01912*.

Stefanie Wulff. 2008. *Rethinking Idiomaticity: A Usage-based Approach*. A&C Black.

Kayo Yin and Graham Neubig. 2022. Interpreting Language Models with Contrastive Explanations. In *Proceedings of EMNLP*.

Appendix

A GPT2 parameters

	layers	hidden states	heads	parameters
GPT2	12	768	12	110M
GPT2-medium	24	1024	16	345M
GPT2-large	36	1280	20	774M
GPT2-xl	48	1600	25	1558M

Table 3: Details of GPT2 model parameters.