# Investigating the Effect of Discourse Connectives on Transformer Surprisal: Language Models Understand Connectives; *Even So* They Are Surprised

**Yan Cong**
Purdue University
cong4@purdue.edu

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

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

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

## Abstract

As neural language models (NLMs) based on Transformers are becoming increasingly dominant in natural language processing, several studies have proposed analyzing the semantic and pragmatic abilities of such models.

In our study, we aimed at investigating the effect of discourse connectives on NLMs with regard to Surprisal scores. We did this by focusing on the English stimuli of an experimental dataset, in which the expectations about an event in a discourse fragment could be reversed by a concessive or a contrastive connective.

By comparing the Surprisal scores of several NLMs, we found that bigger NLMs show patterns similar to humans' behavioral data when a concessive connective is used, while connective-related effects tend to disappear with a contrastive one. We have additionally validated our findings with GPT-Neo using an extended dataset, and results mostly show a consistent pattern.

## 1 Introduction

Psychologists and cognitive scientists have claimed that understanding a discourse involves constructing a *situation model*; that is, a dynamic mental representation of the state of affairs denoted by the text (Van Dijk and Kintsch, 1983; Zwaan and Radvansky, 1998). Extensive evidence has shown that humans use a general knowledge of events and their connections to anticipate upcoming input in the process of language comprehension (McRae and Matsuki, 2009). In this sense, *discourse connectives* might play an important role in updating human situation models and in modulating our expectations about "what is coming next", because concessive (such as *even so*) and contrastive connectives (such as *however*) signal to the comprehender that the upcoming proposition is going to contradict what was previously said, or negate the previous expectations (Xiang and Kuperberg, 2015).

Experimental studies have shown that such connectives have a facilitating effect on human sentence processing (Asr and Demberg, 2020), especially when humans are processing incoherent words and scenarios (Xiang and Kuperberg, 2015). Consider the following example taken from Xiang and Kuperberg (2015), in which the concessive connective *even so* causes an effect of *expectation reversal*:

(1) Liz took the test and failed it. **Even so**, she went home and <u>celebrated</u> wildly.

Given the scenario described in the first sentence (failing a test), the underlined verb in the second sentence is surprising and unexpected. However, after including a connective reversing the readers' expectations, examples like (1) are considered as coherent by human speakers.

The recent literature on natural language processing (NLP) has shown an increasing interest in the use of **Surprisal** scores (Hale, 2001; Levy, 2008) computed by neural language models (NLMs) to account for sentence processing phenomena (Futrell et al., 2018; Van Schijndel and Linzen, 2018; Wilcox et al., 2018), including facilitation (Michaelov and Bergen, 2020, 2022a,b; Michaelov et al., 2023) and interference effects (Cong et al., 2023) in online sentence processing. However, to the best of our knowledge, no studies have attempted to model the facilitation effects of concessive and contrastive connectives at different levels of discourse coherence thus far.

In our study, we aim to fill this research gap by investigating the effect of discourse connectives on NLMs' Surprisal scores. First, we focus on the concessive connective *even so*, and on the contrastive connective *however* as an alternative.

Based on the whole discourse, we first computed the Surprisal scores for target words using NLMs to observe the extent to which they were affected by

*Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 222–232
December 7, 2023. ©2023 Association for Computational Linguistics

the coherence of the stories. We found that NLMs, and particularly the larger models, show patterns that are quite similar to human behavioral data. Moreover, we noticed that the connective-related effects do not show up with contrastive connective, suggesting that the NLMs are sensitive to the difference between connective types: the semantics of concessive connectives entails a reversal of previous expectations about an upcoming event that is not conveyed by contrastive connectives. Using our biggest model, GPT-Neo, we ran additional analysis adding more connectives of the two types and computing the Surprisal scores either in an inter-sentential and an intra-sentential setting. The results were mostly consistent with our first experiment, corroborating the previous findings.

## 2 Related Work

### 2.1 NLM Estimation of Word Surprisal

Transformer-based NLMs (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2019) have become increasingly popular in NLP research, and a number of studies designed tests to investigate their actual linguistic abilities (Tenney et al., 2019a; Jawahar et al., 2019; Tenney et al., 2019b). Some of these studies have specifically analyzed the **Surprisal** scores computed by the models, to understand the extent to which they are sensitive to linguistic phenomena that have been shown to affect human sentence processing. For example, Misra et al. (2020) investigated the predictions of BERT in a setting aimed at reproducing human semantic priming; they reported that BERT was indeed sensitive to "priming", and predicted a word with lower Surprisal values when the context included a related word as opposed to an unrelated one. Using a similar methodology, Cho et al. (2021) modeled the priming effect of verb aspect on the prediction of typical event locations, finding that BERT outputs lower Surprisal scores for typical locations. However, differently from humans, it does so regardless of verb aspect.

Michaelov and Bergen (2022a) investigated the issue of collateral facilitation; that is, a scenario when anomalous words in a sentence are processed more easily by humans due to the presence of semantically related words in the context. They compared the Surprisal scores obtained from several Transformers NLMs and found that most of them reproduced the same significant differences between the conditions that were observed by humans' behaviors. Michaelov et al. (2023) used NLM surprisal scores to replicate the effect of the discourse context in reducing the N400 amplitude for anomalous words, using the Dutch stimuli in the experiments by Nieuwland and Van Berkum (2006).[1]

### 2.2 Discourse Connectives in NLP

The importance of connectives in NLP research is due to the fact that they lexicalize specific discourse relations (Braud and Denis, 2016; Ma et al., 2019). During the acquisition of annotations for discourse-parsing tasks, the connectives sometimes provide a clue to the discourse relations, which are sometimes implicit. In such cases, human annotators are asked to insert the connective that they consider to be more appropriate (Yung et al., 2019).

Ko and Li (2020) proposed to investigate GPT-2's linguistic competence in terms of discourse coherence by testing the model's ability to produce the correct connectives, when given a discourse relation linking two clauses. Using both organic generation and fine-tuned scenarios, they observed that GPT-2 did not always generate coherent discourse, although the generations were better aligned with human behavior in the fine-tuned scenario.

Pandia et al. (2021) evaluated several NLMs on the prediction of the correct connectives in contexts that required Gricean-like pragmatic knowledge and in which a specific connective would correspond to an implicature. For example, in cases such as *Maggie did the paperwork by hand* ***and*** *the company bought new computers, which is to say, Maggie did the paperwork by hand [MASK] the company bought new computers.*, the model had to predict *before* in the [MASK] position to show an understanding that the implied meaning of *and* in this context was *and then*). The authors showed that, when controlling strictly for low-level lexical and syntactic cues, the models performed at chance level at best.

In contrast to previous studies, we did not ask the NLMs to predict a missing connective in a cloze

[1] The N400 is one of the most widely studied component in the literature on event-related potentials (ERP). The N400 component is a negative-going deflection that peaks around 400 milliseconds after presentation of the stimulus word and, although there are different interpretations of its meaning, there is a general agreement among researchers that it may represent a sort of brain signature of semantic complexity (Hagoort, 2003). Therefore, a *reduced* N400 amplitude due to the presence of semantically-related words in the discourse context can be interpreted as a facilitation effect.

setting; instead, we analyzed the impact of a concessive/contrastive connective on the model's expectations for a given event, which might be coherent or not with the scenario. In practical terms, this translates into analyzing the Surprisal of the model at the verb in the subordinate clause: we predict that if the model is linguistically competent and can identify coherence correctly, the coherent items should be assigned *lower* Surprisal scores.

| Scenario type | Cloze probability | Coherence |
|---|---|---|
| Coherent | 0.42 | 4.8 |
| Incoherent | 0.03 | 1.7 |
| Even-so Coherent | 0.31 | 3.3 |
| Even-so Incoherent | 0.04 | 2.4 |

Table 1: Summary table for the human data in Xiang and Kuperberg (2015). Cloze probability is represented as the proportion of total responses from 40 participants. 5: very coherent; 1: incoherent.

## 3 Experimental Settings

### 3.1 Dataset

We used the English stimuli provided by Xiang and Kuperberg (2015), who designed 180 sets of two-sentence discourse items, each with four conditions as in (2) (45 scenarios per condition). The target word (underlined) appeared in the final sentence.

(2) a. Liz had a history exam on Monday. She took the test and **aced** it. She went home and *celebrated* wildly. (*Plain*, *Coherent*)

b. Liz had a history exam on Monday. She took the test and **failed** it. She went home and *celebrated* wildly. (*Plain*, *Incoherent*)

c. Liz had a history exam on Monday. She took the test and **failed** it. *Even so*, she went home and *celebrated* wildly. (*Even so*, *Coherent*)

d. Liz had a history exam on Monday. She took the test and **aced** it. *Even so*, she went home and *celebrated* wildly. (*Even so*, *Incoherent*)

We also created alternative versions of (2-c) and (2-d) by replacing *Even so* with *However*. Note that, as *however* is a contrastive connective, its semantics signals an upcoming contrast, but not necessarily the denial of previously-held expectations as in concessive relations (Izutsu, 2008), and thus, it was interesting for us to test and compare the consistency of the reversal effect.

Xiang and Kuperberg (2015) collected cloze probabilities and typicality judgments for their items (Table 1). The coherent items had the highest cloze probability scores and coherence ratings, whereas the incoherent items had the lowest ones. The coherent *even-so* items exhibited significantly lower cloze probability and coherence ratings than the plain coherent ones did; while the incoherent *even-so* items were rated as more plausible than the plain incoherent ones, the difference was not significant.

Their EEG experiment showed some differences from the behavioral data: The N400 component for the target verb was more reduced in the coherent *even-so* items (i.e., lower processing costs), compared to the plain coherent ones, while incoherent items with *even-so* showed higher processing costs than the plain incoherent ones at the target verb, eliciting a P600 component.[2]

### 3.2 Language Models

For the models in this paper, we use the implementation of Minicons (Misra, 2022)[3], which is an open source library that provides a standard API for behavioral and representational analyses of NLMs. We experimented with three variants of autoregressive LMs of different sizes: the original GPT-2 Base, with 124 million parameters (Radford et al., 2019); DistilGPT-2 with 82 million parameters (Sanh et al., 2019), which was trained as a student network with the supervision of GPT-2; and our biggest model GPT-Neo, with 1.3 billion parameters (Gao et al., 2020; Black et al., 2021).

Using autoregressive NLMs, we computed the Surprisal scores for the targets in the stimuli - the critical verb in the final clause. Notice that, in the four conditions of the same item, the verb to be predicted is always the same. More formally, the Surprisal for the target $T$ in the context $C$ (**Surp**) was computed as:

$$Surp(w_t) = -logP(w_t|w_{1...t-1}) \tag{1}$$

When $w_t$ was tokenized into multiple subword tokens, we simply used the average of the subword

[2] The P600 is positive-going wave peaking around 600 ms after the presentation of a stimulus word. In online sentence processing studies, it is generally associated with the presence of syntactic anomalies and structural reprocessing (Osterhout and Holcomb, 1993; Luck, 2014).

[3] `https://github.com/kanishkamisra/minicons-experiments`

| | GPT-2 | DistilGPT-2 | GPT-Neo |
|---|---|---|---|
| Intercept | *** | *** | *** |
| DisCohere | *** | *** | *** |
| DisConn | ** | *** | *** |
| length | | | * |
| DisCohere:<br>DisConn | *** | *** | *** |

Table 2: *Even so* dataset: Summary table for the significance scores of different predictors of **Surp**. Notation: $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$.

| | GPT-2 | DistilGPT-2 | GPT-Neo |
|---|---|---|---|
| Intercept | *** | *** | *** |
| DisCohere | *** | *** | *** |
| DisConn | | | |
| length | | | * |
| DisCohere:<br>DisConn | * | | *** |

Table 3: *However* dataset: Summary table for significance scores of the different predictors of **Surp**.

| | GPT-2 | DistilGPT-2 | GPT-Neo |
|---|---|---|---|
| CohereNoconn<br>CohereConn | $-0.45^{*}$ | $-0.426^{*}$ | $-0.89^{***}$ |
| IncohereNoconn<br>IncohereConn | $0.365^{*}$ | $0.475^{***}$ | $0.716^{***}$ |
| IncohereConn<br>CohereConn | 0.23 | -0.205 | $0.539^{**}$ |
| IncohereNoconn<br>CohereNoconn | $1.044^{***}$ | $0.695^{***}$ | $2.144^{***}$ |

Table 4: *Even so* dataset: Summary table for the *estimate* and the *p*-values of the pairwise comparisons.

| | GPT-2 | DistilGPT-2 | GPT-Neo |
|---|---|---|---|
| CohereNoconn<br>CohereConn | $-0.49^{*}$ | -0.211 | $-1.228^{***}$ |
| IncohereNoconn<br>IncohereConn | -0.151 | 0.047 | 0.066 |
| IncohereConn<br>CohereConn | $0.715^{***}$ | $0.455^{***}$ | $0.857^{***}$ |
| IncohereNoconn<br>CohereNoconn | $1.054^{***}$ | $0.712^{***}$ | $2.15^{***}$ |

Table 5: *However* dataset: Summary table for the *estimate* and *p*-values of the pairwise comparisons.

tokens probabilities.[4] However, we found that this happens only for the 14% of the target verbs in the dataset (only 23 out of 163 targets are not included in the models' vocabulary).

For each NLM, we fitted a linear mixed-effects model using the Surprisals (**Surp**) of the target verbs as the dependent variable. The independent variables include: the coherence of the discourse *DisCohere* (coherent vs. incoherent), the presence of discourse connectives *DisConn* (with connective vs. plain/without connective), their interaction (*DisCohere:DisConn*), and the token length of the stimulus (*length*). We used the ID of the items (ITEM_ID) as the random intercept in our models. We used the *lmerTest* package (Kuznetsova et al., 2017) for model fitting and results; finally, the pairwise comparisons with Tukey adjustment were carried out by means of the EMMEANS package (Lenth, 2019) in R.

## 4 Results

For the original *Even so* data (Table 2), our results revealed that all three NLMs showed significant sensitivity to the coherence of the discourse (*DisCohere*) and to the presence of connectives (*DisConn*). Interaction effects were found in all the NLMs, and only GPT-Neo showed effects on length. Interestingly, the replacement of *Even so* with *However* caused the *DisConn* effects to disappear in all the NLMs (Table 3). Interaction effects were found in GPT-2 and GPT-Neo, and again, only GPT-Neo showed sensitivity to length.

The pairwise comparisons examining the effects of *Even so* at each level of Coherence (Table 4) showed that, for all the models, there is a decrease of Surprisals from *Even so* coherent scenarios (condition c.) to plain coherent scenarios (condition a.), and an increase of Surprisals from *Even so* incoherent scenarios (condition d.) to plain incoherent scenarios (condition b.). Pairwise comparisons examining effects of Coherence at each level of *Even so* showed a significant increase of Surprisals from *Even so* coherent scenarios (condition c.) to *Even so* incoherent scenarios (condition d.) only with GPT-Neo. All the NLMs showed an increase of Surprisals from plain coherent scenarios (condition a.) to plain incoherent ones (condition b.).

Fewer significant effects were found after replacing *Even so* with *However* (Table 5). Regarding the effects of *However* at each level of Coherence, GPT-2 and GPT-Neo revealed a decrease of Surprisals from *However* coherent scenarios (condition c.) to plain coherent scenarios (condition a.). All NLMs showed no significant effects of Surprisals from *However* incoherent condition (condition d.) to plain incoherent condition (condition b.). As for the effects of Coherence at each level of *However*, all the NLMs showed an increase of Surprisals from *However* coherent condition (condition c.) to *However* incoherent condition (condition d.), and an increase of Surprisals from plain

[4] Upon request of the reviewers, results for the experiment with the sum of the Surprisal scores instead of the average can be found in the Appendix.

| | GPT-2 | DistilGPT-2 | GPT-Neo |
|---|---|---|---|
| DisCohere | ✓ | ✓ | ✓ |
| DisConn | ✓ | ✓ | ✓ |
| **DisCohere**: at each level of DisConn | | | ✓ |
| **DisConn**: at each level of DisCohere | ✓ | ✓ | ✓ |

Table 6: *Even so* dataset: Comparison of effects between Human behavioral results and NLMs Surprisals. Notation: ✓ = alignment with Human in significance and direction of the effect.

coherent condition (condition a.) to plain incoherent condition (condition b.).

Comparing the outcome with the study by Xiang and Kuperberg (2015), one can observe that the model scores tend to align with human typicality judgements (cf. Table 1), and the largest one (i.e. GPT-Neo) shows the same effect pattern (cf. Table 6). A difference, however, is that all the NLMs assign significantly higher Surprisals to plain incoherent items than *Even so* incoherent ones.

Our results suggest that NLMs are sensitive to the expectation reversal determined by connectives. Besides the human-like pattern in the distribution of the Surprisal scores for the *Even so* dataset, it is also noticeable that replacing the connective with *However* makes the connective-related effects disappear. This is coherent with the intuition and the claims made in formal semantics literature, for which *However* simply introduces a semantic opposition, while *Even so* additionally presupposes an expectation being denied (Karttunen and Peters, 1979; Izutsu, 2008).

### 4.1 Extended Study

Our experiments on Surprisal suggest that our larger NLM, GPT-Neo, shows similar patterns to humans behavioral data with the concessive connective *even so*. Interestingly, all NLMs show distinct patterns with concessive and contrastive connective, with no connective-related effects when *however* is used. This might be due to the fact that contrastive connectives *per se* just indicate a semantic opposition, but differently from concessive ones, they do not necessarily deny an expectation about an event. However, one might ask if the NLMs would consistently score the discourse items even when using different concessive or contrastive connectives.

To verify this, we extended our study with GPT-Neo: 1) we selected more connectives for the two groups, a) *but, yet* and *still* for the contrastive group and b) *nonetheless, nevertheless* and *regardless* for the concessive one; in each item of the original stimuli by Xiang and Kuperberg (2015), we replaced the original *even so* connectives with the new ones, obtaining 6 new datasets (one for each of the newly-introduced connectives); 2) NLMs predictions have been shown to be extremely sensitive even to small changes in the input (Jiang et al., 2020); in our case, the predictions might have been affected by the fact that the connectives always appeared in a new sentence after a full stop (**inter-sentential** setting). Therefore, we also carry out the experiment after replacing the final full stop of the second sentence with a comma (**intra-sentential** setting), and lower-casing the discourse connectives, as it can be seen in Example (3):

(3) a. Liz took the test and failed it. Even so, she went home and celebrated wildly. (*inter-sentential*)
b. Liz took the test and failed it, even so, she went home and celebrated wildly. (*intra-sentential*)

Our choice of connectives was based on Webber et al. (2019), which describes and annotates *but, yet, still* as contrastive connectives that share the same syntax and semantics as *however*, and *nonetheless, nevertheless, regardless* as concessive connectives that introduce events in the same manner as *even so* does. Moreover, using those connectives it was easy to modify our stimuli by replacing the original *even so* and maintaining at the same time the same word order and syntax of the experimental items.

The procedure for computing the Surprisal scores with the NLMs and running the linear mixed models is the same of Section 3.2, but this time we only used GPT-Neo, as it was the model with the most similar pattern to human behavioral data.

As shown in Tables 7-8, the results suggest that the pattern found in inter-sentential *even so/however* mostly gets reproduced across different inter-/intra-sentential connectives. *still* is the connective that shows more discrepancy: as intra-sentential connective, we found connectives (DisConn) effects, which were absent in *however*. It is interesting to notice that the presence or not of the DisConn effect is what sets apart the two sets of connectives: similarly to *even so*, *nevertheless, nonetheless* and *regardless* all display significant

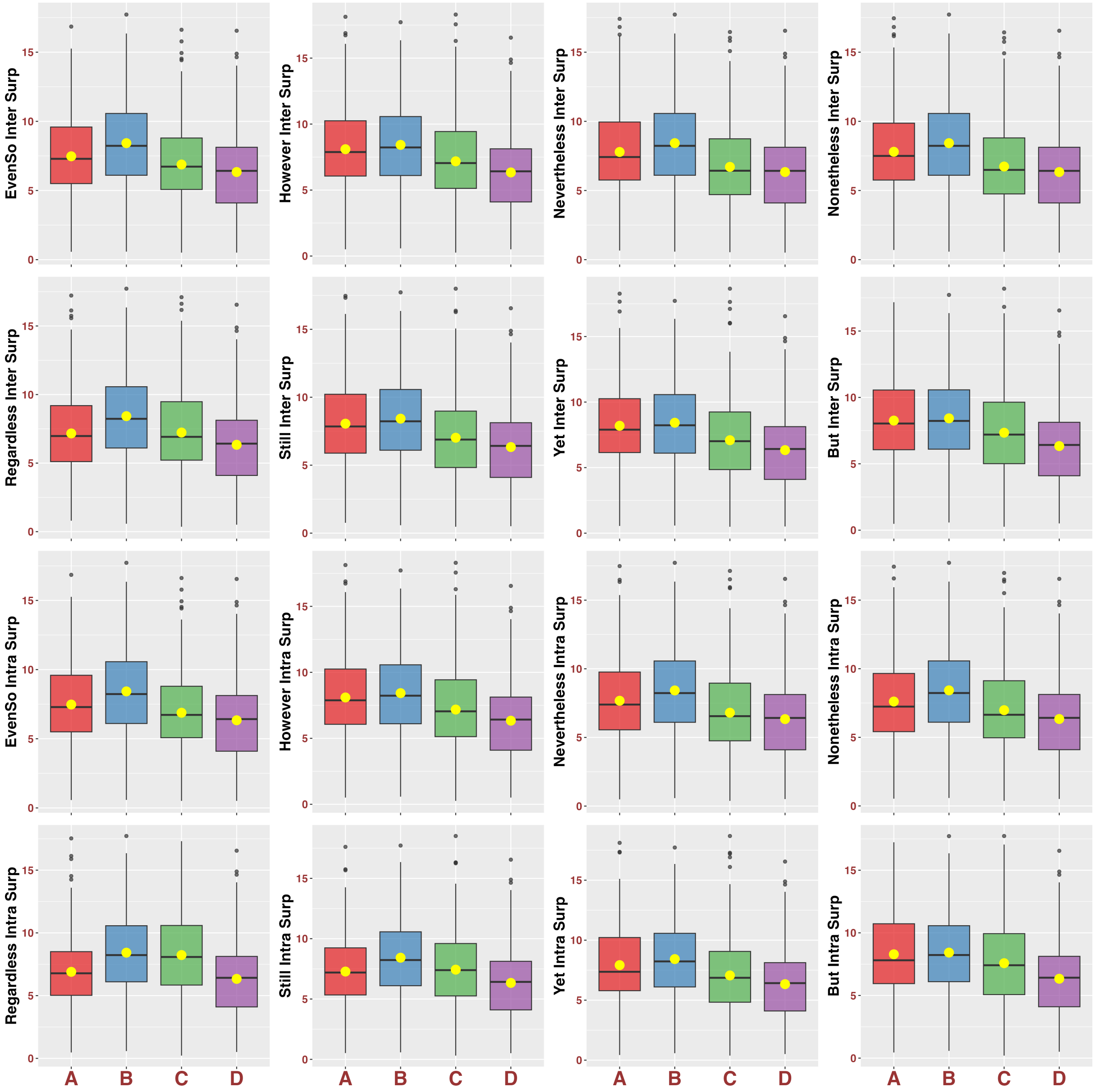

Figure 1: Boxplots of the GPT-Neo Surprisals for all the inter-/intra-sentential connectives (mean of the scores is marked in yellow). Notation: A (red): IMPLAUS_CONN; B (blue): IMPLAUS_NOCONN; C (green): PLAUS_CONN; D (purple): PLAUS_NOCONN.

| | *even so* | *nevertheless* | *nonetheless* | *regardless* | *still* | *yet* | *but* | *however* |
|---|---|---|---|---|---|---|---|---|
| Intercept | *** | *** | *** | *** | *** | *** | *** | *** |
| DisCohere | *** | *** | *** | *** | *** | *** | *** | *** |
| DisConn | *** | * | * | *** | | | | |
| length | * | | * | * | * | * | ** | * |
| DisCohere:<br>DisConn | *** | *** | *** | *** | *** | *** | *** | *** |

Table 7: Extended INTER dataset: Summary table for the significance scores of different predictors of **Surp** using GPT-Neo. Notation: $* = p < 0.05$, $** = p < 0.01$, $* * * = p < 0.001$.

| | *even so* | *nevertheless* | *nonetheless* | *regardless* | *still* | *yet* | *but* | *however* |
|---|---|---|---|---|---|---|---|---|
| Intercept | *** | *** | *** | *** | *** | *** | *** | *** |
| DisCohere | *** | *** | *** | *** | *** | *** | *** | *** |
| DisConn | *** | ** | ** | *** | *** | | | |
| length | * | * | * | *** | ** | * | ** | * |
| DisCohere:<br>DisConn | *** | *** | *** | *** | *** | *** | *** | *** |

Table 8: Extended INTRA dataset: Summary table for the significance scores of different predictors of **Surp** using GPT-Neo. Notation: $* = p < 0.05$, $** = p < 0.01$, $* * * = p < 0.001$.

effects, while the contrastive connectives *yet* and *but* do not. *Still* represents the exception to this pattern, and a possible reason might be its ambiguity, as this word can appear as a noun, an adjective, a verb or an adverb, besides its connective usage.

The scores for all settings can be visualized in Figure 1. Across connectives types, GPT-Neo showed the highest Surprisals scores in the incoherent without connectives (IMPLAUS_NOCONN) condition, whereas the lowest scores were observed in the coherent without connectives (PLAUS_NOCONN) condition. We did not find a lot of variance across conditions. We also observed a few outliers, mostly occurring in the coherent with connectives (PLAUS_CONN) condition.

We conducted follow-up comparisons and summarized our results in Table 9. In most cases, for main and interaction effects, GPT-Neo's behavior across concessive connectives aligned well with *even so* in both the statistical significance and direction of the effects. A discrepancy was found in the follow-up comparisons: the inter-/intra-sentential *regardless* did not align well with *even so* for the coherence effects with respect to connectives (DisCohere: at each level of DisConn).

Similarly to *still*, we speculate that a possible reason could be the ambiguity of this word, as *regardless* can appear in a sentence as an adjective or as a preposition (with the meaning of *in spite of/despite*), and thus it might lead the NLM to less accurate predictions. Interestingly, and differently from the other concessive, it can be noticed from Figure 1 (in the first boxplots of the second and of the fourth row) that with the *regardless* connective the IMPLAUS_CONN items (red boxes) tend to have similar, or lower Surprisal scores than the PLAUS_CONN ones (green boxes).

Concerning contrastive connectives, inter-/intra-sentential *but* did not align with *however* for the connectives effects with respect to coherence (DisConn: at each level of DisCohere). Additionally, our findings indicate that intra-sentential *still* did not align with *however* regarding coherence effects (DisCohere: at each level of DisConn), and that inter-sentential *yet* did not align with *however* regarding connective effects (DisConn: at each level of DisCohere). In general, contrastive connectives are less consistent with regard to the pattern found in the original experiment, showing that, for how the stimuli were built, the denial of expectations introduced by a concessive connective is an important cue for modulating the coherence of the continuation of the story.

## 5 Conclusion

In our paper, we proposed an analysis of the Surprisal scores of NLMs on the target verbs of a psycholinguistic dataset where the items differed by the coherence of the discourse, and by the inclusion of a connective reversing the expectations on the verb. We found that our NLMs show patterns that are quite similar to human behavioral data, in particular the biggest model, GPT-Neo. More interestingly, in all models the effects related to the connective disappear when a contrastive connective is used to replace the concessive one. This

| | *nevertheless* | *nonetheless* | *regardless* | *still* | *yet* | *but* |
|---|---|---|---|---|---|---|
| | Align with inter-sentential *even so* | | | Align with inter-sentential *however* | | |
| DisCohere | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| DisConn | ✓ | ✓ | ✓ | | ✓ | ✓ |
| **DisCohere**: at each level of DisConn | ✓ | ✓ | | ✓ | ✓ | ✓ |
| **DisConn**: at each level of DisCohere | ✓ | ✓ | ✓ | ✓ | | |
| | Align with intra-sentential *even so* | | | Align with intra-sentential *however* | | |
| DisCohere | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| DisConn | ✓ | ✓ | ✓ | | ✓ | ✓ |
| **DisCohere**: at each level of DisConn | ✓ | ✓ | | | ✓ | ✓ |
| **DisConn**: at each level of DisCohere | ✓ | ✓ | ✓ | ✓ | ✓ | |

Table 9: Extended dataset: comparing GPT-Neo Surprisal scores across connectives. Notation: ✓ = aligned in significance and direction of the effect.

suggests that the concessive connective leads to an expectation reversal for the upcoming verb, while the contrastive one does not, coherently with previous descriptions of connectives from the formal semantics literature (Karttunen and Peters, 1979; Izutsu, 2008).

Given that the psycholinguistic dataset we used in our modeling is relatively small for NLP settings and limited to two connectives, we additionally constructed an extended datasets to validate our findings. We extended our investigation in two ways: expanding the dataset by varying the setting in which the connective was found (inter- or intra-sentential) and connectives themselves, including six more contrastive/concessive connectives. Our results indicate that the findings *even so* are mostly consitent, as they generalize across settings and concessive connectives.

We acknowledge that there are some significant limitations in our current investigations. First, we have human data only for *even so* in the inter-sentential setting, for which we could establish an interpretation baseline. There is no behavioral or neural data for all the other connectives. This implies that we interpreted some of our findings based on intuitions about discourse connectives, assuming that the human behavioral pattern will be similar for other concessive types. We recognize that this is an important limitation to the cognitive plausibility of our evaluation, and for future research we plan to collect more human judgements for discourse connectives, possibly including also languages other English.

Second, our analysis was mainly focused on comparing Surprisal scores to human behavioral patterns (or, by extension, to the patterns found in the original experiment with concessive connectives), but we did not apply any advanced interpretability method to identify which specific input tokens influence the predictions for the target verb. More direct evidence for the causal role of discourse connectives in reversing the predictions could be obtained, for example, by analyzing the changes of the probability rank of the verb in the target position; or by applying contrastive explanations to sentences differing only for the presence of connectives (Yin and Neubig, 2022).

## Acknowledgements

This research was funded by the PROCORE grant at the Hong Kong Polytechnic University (project 3-RABX, "Exploring the Influence of Discourse Connectives on the Predictions of Humans and Pre-trained Language Models"). We would also like to thank their three anonymous reviewers for their constructive feedback.

## References

Fatemeh Torabi Asr and Vera Demberg. 2020. Interpretation of Discourse Connectives Is Probabilistic: Evidence from the Study of But and Although. *Discourse Processes*, 57(4):376–399.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. *zenodo.org*.

Chloé Braud and Pascal Denis. 2016. Learning Connective-based Word representations for Implicit Discourse Relation Identification. In *Proceedings of EMNLP*.

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*.

Yan Cong, Emmanuele Chersoni, Yu-Yin Hsu, and Alessandro Lenci. 2023. Are Language Models Sensitive to Semantic Attraction? A Study on Surprisal. In *Proceedings of *SEM*.

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*.

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

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027*.

Peter Hagoort. 2003. Interplay Between Syntax and Semantics During Sentence Comprehension: ERP Effects of Combining Syntactic and Semantic Violations. *Journal of Cognitive Neuroscience*, 15(6):883–899.

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

Mitsuko Narita Izutsu. 2008. Contrast, Concessive, and Corrective: Toward a Comprehensive Study of Opposition Relations. *Journal of Pragmatics*, 40(4):646–675.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of ACL*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Lauri Karttunen and Stanley Peters. 1979. Conventional Implicature. In *Presupposition*, pages 1–56. Brill.

Wei-Jen Ko and Junyi Jessy Li. 2020. Assessing Discourse Relations in Language Generation from GPT-2. In *Proceedings of INLG*.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13):1–26.

Russell Lenth. 2019. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R Package Version 1.4.2.

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

Steven J Luck. 2014. *An Introduction to the Event-related Potential Technique*. MIT Press.

Mingyu Derek Ma, Kevin K Bowden, Jiaqi Wu, Wen Cui, and Marilyn Walker. 2019. Implicit Discourse Relation Identification for Open-domain Dialogues. In *Proceedings of ACL*.

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.

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*.

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*.

Mante S Nieuwland and Jos JA Van Berkum. 2006. When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. *Journal of Cognitive Neuroscience*, 18(7):1098–1111.

Lee Osterhout and Phillip J Holcomb. 1993. Event-related Potentials and Syntactic Anomaly: Evidence of Anomaly Detection During the Perception of Continuous Speech. *Language and Cognitive Processes*, 8(4):413–437.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic Competence of Pre-trained Language Models through the Lens of Discourse Connectives. In *Proceedings of CONLL*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. In *Open-AI Blog*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. In *Proceeding of the NeurIPS EMC*$^2$$Workshop$.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of ACL*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019b. What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations. *arXiv preprint arXiv:1905.06316*.

Teun Adrianus Van Dijk and Walter Kintsch. 1983. *Strategies of Discourse Comprehension*. Academic Press New York.

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.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual. *Philadelphia, University of Pennsylvania*, 35:108.

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*.

Ming Xiang and Gina Kuperberg. 2015. Reversing Expectations during Discourse Comprehension. *Language, Cognition and Neuroscience*, 30(6):648–672.

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

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing Discourse Relation Annotations by a Two-step Connective Insertion Task. In *Proceedings of the ACL Linguistic Annotation Workshop*.

Rolf A Zwaan and Gabriel A Radvansky. 1998. Situation Models in Language Comprehension and Memory. *Psychological Bulletin*, 123(2):162.

# A Appendix

Upon request of the reviewers, we additionally rerun the experiment by computing the sum of the Surprisal scores of the subtokens (**Surp-sum**) for out-of-vocabulary target verbs, instead of taking the average.The findings are summarized in Tables 10-11.

The results mostly show consistent patterns. Compared with the average of the surprisals of the subtokens (Tables 7-8), with the sum the DisConn effect in inter-sentential *nevertheless* disappears, while weakly-significant DisConn effects appear for the inter-sentential *still* and for the intra-sententials *yet* and *however*.

| | *even so* | *nevertheless* | *nonetheless* | *regardless* | *still* | *yet* | *but* | *however* |
|---|---|---|---|---|---|---|---|---|
| Intercept | *** | *** | *** | *** | *** | *** | *** | *** |
| DisCohere | *** | *** | *** | *** | *** | *** | *** | *** |
| DisConn | *** | | ** | ** | *** | | | |
| length | | * | | | | | | * |
| DisCohere: DisConn | *** | *** | *** | *** | *** | *** | *** | *** |

Table 10: Extended INTER dataset: Summary table for the significance scores of different predictors of **Surp-sum** using GPT-Neo. Notation: $* = p < 0.05$, $** = p < 0.01$, $* * * = p < 0.001$.

| | *even so* | *nevertheless* | *nonetheless* | *regardless* | *still* | *yet* | *but* | *however* |
|---|---|---|---|---|---|---|---|---|
| Intercept | *** | *** | *** | *** | *** | *** | *** | *** |
| DisCohere | *** | *** | *** | *** | *** | *** | *** | *** |
| DisConn | *** | *** | *** | *** | *** | * | | * |
| length | * | | | ** | ** | | ** | * |
| DisCohere: DisConn | *** | *** | *** | *** | *** | *** | *** | *** |

Table 11: Extended INTRA dataset: Summary table for the significance scores of different predictors of **Surp-sum** using GPT-Neo. Notation: $* = p < 0.05$, $** = p < 0.01$, $* * * = p < 0.001$.