Otherwise in Context: Exploring Discourse Functions with Language Models

Guifu Liu[1] and **Bonnie Webber**[1] and **Hannah Rohde**[2]
[1]School of Informatics, University of Edinburgh
[2]School of Philosophy, Psychology and Language Sciences, University of Edinburgh
Guifu.Liu@uni-saarland.de {Bonnie.Webber, Hannah.Rohde}@ed.ac.uk

Abstract

Discourse adverbials are key features of discourse coherence, but their function is often ambiguous. In this work, we investigate how the discourse function of *otherwise* varies in different contexts. We revise the function set in Rohde et al. (2018b) to account for a new meaning we have encountered. In turn, we create the *otherwise* corpus, a dataset of naturally occurring passages annotated for discourse functions, and identify lexical signals that make a function available with a corpus study. We define *continuation acceptability*, a metric based on surprisal to probe language models for what they take the function of *otherwise* to be in a given context. Our experiments show that one can improve function inference by focusing solely on tokens up to and including the head verb of the continuation (i.e., *otherwise* clause) that have the most varied surprisal across function-disambiguating discourse markers. Lastly, we observe that some of these tokens confirm lexical signals we found in our earlier corpus study, which provides some promising evidence to motivate future pragmatic studies in language models.[1]

1 Introduction

Discourse coherence helps us understand what a speaker or writer is trying to say in placing one segment of text next to another (Kehler, 2006). In this paper, we focus on a key aspect of discourse coherence: the discourse adverbial *otherwise*, a word whose function in discourse depends on both its lexical semantics and a pragmatic understanding of the context. As seen in Figure 1, *otherwise* can convey 1) CONSEQUENCE: what would happen when a situation doesn't occur, 2) ENUMERATION: what is another option to achieve some goal, and 3) EXCEPTION: what is usually the case given that the clause left of *otherwise*, or left hand side [LHS] conveys an exception.

[1]Code and data are available in https://github.com/GuifuLiu/otherwise

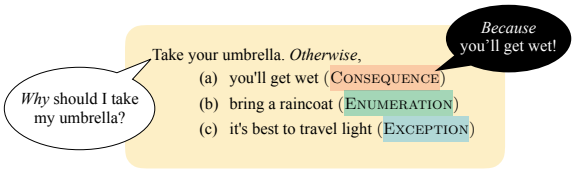

Figure 1: An example of *otherwise* functions.

Being able to distinguish these discourse functions is important for downstream applications in natural language understanding. For example, when asked "Why should I take an umbrella?", a question-answering system should apply clause (a) and respond with "to avoid getting wet", rather than clause (b) "to avoid bringing a raincoat".

Although Rohde et al. (2018b) have shown that human participants can distinguish these discourse functions in a provided context, we still don't understand the signals that make such a function available, and the extent to which language models can infer these discourse functions. In addition, previous work has been limited to small-scale, researcher-constructed examples in the context of psycholinguistic studies (Rohde et al., 2016, 2018a).

To shine light on these questions, we introduce the *otherwise* corpus, a dataset of 294 naturally-occurring passages annotated for discourse functions, and a revised *otherwise* function set, to account for a new meaning that is not discussed in Rohde et al. (2018b). Through corpus study, we find that these respective functions are associated with the presence of distinctive lexical cues such as negation markers, modal triggers, and conjunctions.

To study how language models infer the function of *otherwise* in context, we define *continuation acceptability*: we replace *otherwise* with a set of candidate discourse markers that are distinctive of a function (e.g., *alternatively* for ENUMERATION). We expect that the model will accept the one that

best fits the context by assigning low *surprisal* (a word's negative log probability in context) to the *continuation* (the text segment after a candidate marker, or right hand side [RHS]). We validate this metric by showing that it can infer the annotators' assigned function better than a majority baseline, though its ability to do so varies across discourse functions of the passage and models.

We then explore alternative aggregation methods beyond the per-token average used in *continuation acceptability* to identify key tokens that help signal the function of *otherwise*. We find that solely focusing on tokens up to and including the head verb of the continuation that have the most varied surprisal across discourse markers shows convincing performance improvement, despite ignoring other tokens. In addition, some of these tokens confirm lexical signals identified in the corpus study, suggesting that when the model infers the *otherwise* function, these signals are indeed relevant.

Our contributions are (i) the *otherwise* corpus, a dataset of naturally-occurring *otherwise* passages annotated for discourse functions (§3.2), (ii) *continuation acceptability*, a new metric based on language models to probe for their most accepted discourse function (§3.4), (iii) insights into how lexical signals help make a discourse function available (§3.3), and (iv) results showing how language models are affected by certain aspects of the context (§4.2, §4.3).

2 Related Work

Theories of discourse coherence shape our research questions and inform our experimental design. In what follows, we begin by outlining prior work on interpreting *otherwise* in context (§2.1). We then discuss the application of language models in discourse research (§2.2).

2.1 *Otherwise* in Context

Knott (1996) studied the semantics of *otherwise* in relation to other discourse markers with a substitution test to discover when a writer is willing to substitute *otherwise* for another discourse marker. *Otherwise* was found to be synonymous with *if not*, a hyponym of *or* and *or else*, and contingently substitutable with *alternatively*. The finding suggests that *otherwise* exhibits granularity in its semantic meaning.

Webber et al. (1999) noted that *otherwise* is compatible with additional discourse relations, such as an unmarked *because* in "If the light is red, stop. Otherwise, you might get run over". Likewise, Rohde et al. (2016) have shown that, in the presence of *otherwise*, people infer additional discourse relations that hold jointly with those associated with the adverbial, by inserting connectives *because*, *or*, *but* before *otherwise*. For instance, in Figure 1, one may insert *because* to indicate inference of ARGUMENTATION for (a), *or* to indicate inference of ENUMERATION for (b), and *but* to indicate inference of EXCEPTION for (c).

Rohde et al. (2018b) subsequently provide empirical evidence for why conjunctions inserted before *otherwise* split among these three. They have found that variability in the choice of conjunctions arises from the lexical semantics of *otherwise*, combined with inferences of its discourse function (to be discussed in §3.1).

Our work builds on previous findings by scrutinizing the lexical signals that make a function available, using large-scale, naturally occurring examples that represent how a speaker or writer uses the discourse adverbial, and examining how language models infer the functions of *otherwise*.

2.2 Discourse and Language Models

The use of language models in discourse is an active research area. Recent work on discourse markers and language models has taken two main approaches: (i) using **cloze tasks** with masked language models to predict connectives (Kurfalı and Östling, 2021; Pandia et al., 2021; Stodden et al., 2023; Dong et al., 2024), and (ii) using **prompting** to insert discourse connectives for implicit discourse relation annotation (Yung et al., 2024) and to uncover the function of discourse particle *actually* (Sadlier-Brown et al., 2024) and *just* (Sheffield et al.).

While standard masked language models may be limited in predicting multi-token discourse markers without additional training (Kalinsky et al., 2023), prompting also has several shortcomings. In particular, small variations in the prompt are shown to affect model outputs (Salinas and Morstatter, 2024; Mizrahi et al., 2024). To avoid the drawbacks of prompting, we use surprisal scores of language models to infer the discourse function of *otherwise*. There is also an increasing interest in the use of surprisal to account for a wide range of linguistic phenomena, such as sentence processing (Wilcox et al., 2018), utterance predictability (Giulianelli

et al., 2023), the semantics of generics and quantifiers (Cilleruelo et al., 2025), and discourse structure (Tsipidi et al., 2024). In our study, we apply surprisal to investigate the discourse function of an ambiguous adverbial.

Surprisal has also been used to test the effect of discourse connectives on discourse coherence. Zhou et al. (2010) constructed synthetic passages by inserting a candidate implicit connective between a pair of arguments. A language model is then used to calculate the perplexity of every token in the constructed passage. The connective from the passage with the lowest mean surprisal is chosen as the best implicit connective for the argument pair.

Cong et al. (2023) used controlled psycholinguistic stimuli and calculated the surprisal of a critical word to test the effect of discourse connectives *even so* and *however* on reversing the expectations about an event. Similarly, we measure how the expectation for the continuation is influenced by candidate discourse markers that disambiguate *otherwise* functions, which may be coherent or not depending on the context. The main differences are that the discourse functions, discourse markers, and the context we investigate are more diverse and complex than those used in psycholinguistic stimuli, which require the model to understand a wider context.

3 Methodology

3.1 Revised Function Set of *Otherwise*

Rohde et al. (2018b) define three functions of *otherwise* based on both the lexical semantics of *otherwise* and the relation that humans infer between two segments in the passage. They are shown in Figure 1. One function is ARGUMENTATION, where the clause to the right of *otherwise*, `[RHS]` shows what the result will be if certain advice in `[LHS]` is not followed, as in (a). Another function is ENUMERATION. When the speaker provides two equally viable options to fulfill a shared goal, `[RHS]` introduces an alternative option, as in (b). A third function is EXCEPTION, where `[RHS]` expresses what is usually the case, while `[LHS]` specifies an exception to it, as in (c).

However, we have encountered an additional meaning of *otherwise* that does not fit into this function set:

> (d) I like you too. Otherwise, we wouldn't be friends.
>
> (e) Of course I mean it. Otherwise, I wouldn't ask.

For these passages, `[RHS]` doesn't provide a reason for the claim in `[LHS]`, an equally viable option, or a description of what generally holds. Instead, the *otherwise* clause describes a logical conclusion if the situation in the `[LHS]` **doesn't** arise. We name this new function CONSEQUENCE.

All ARGUMENTATION passages fulfill the definition of CONSEQUENCE, as their *otherwise* clauses describe an undesirable or negative outcome that can be avoided if the advice in the main clause is followed. However, the opposite is not necessarily true. Therefore, we define ARGUMENTATION as a subordinate function of CONSEQUENCE. However, when we mention CONSEQUENCE as a passage label in the following sections, we refer to passages that are CONSEQUENCE but not ARGUMENTATION.

We provide definitions and examples of the revised function set in Table 1.

3.2 The *otherwise* corpus

The *otherwise* corpus consists of 294 passages with sentence-initial *otherwise* that are annotated for our revised discourse functions in §3.1. These passages are randomly sampled from the *Corpus of Contemporary American English* or COCA (Davies, 2008), and the *British National Corpus* or BNC (BNC Consortium, 2007) and span a wide array of sentence constructions (e.g., declarative, imperative, question), genres and modalities (e.g., blogs, academic, fiction, TV, movies). All passages are contextually contained so that the context provided in the passage is sufficient to infer the discourse function.

To identify the discourse function that is operative in a passage, we use a paraphrase task: for each passage, every function of *otherwise* is assigned a paraphrase to convey the lexical semantics of that function (**Paraphrase** in Table 1), and participants infer a valid paraphrase.

The final dataset contains 294 human-annotated *otherwise* passages and their discourse functions (Table 2). Each passage was annotated by a researcher. In addition, one-fifth of the dataset was also labeled by four participants who are native or near-native adult English speakers. The average inter-annotator agreement between researcher and participant is $\kappa = 0.87$. Details on dataset construction and annotation are in Appendix A.

Function	Definition	Paraphrase	Example
CONSEQUENCE	If the situation in [LHS] doesn't occur, the situation in [RHS] would arise.	[LHS] because if not [LHS], [RHS]	[I like you too.] Otherwise, [we wouldn't be friends.]
↳ ARGUMENTATION	[RHS] is *undesirable* and can be possibly avoided by following [LHS]	To avoid [RHS], [LHS].	[We have to operate immediately.] Otherwise, [she will die.]
ENUMERATION	It doesn't take the failure of [LHS] to consider [RHS] as another option.	There is more than one option for [goal]. They are 1) [LHS] and 2) [RHS].	[I like a nice curry.] Otherwise, [I'll nibble on fruit.]
EXCEPTION	[LHS] is an exception to [RHS]	Generally [RHS], an exception is that [LHS].	[Some people are riding horses.] Otherwise, [people are traveling on foot.]

Table 1: Revised *otherwise* function set, its description, the paraphrase used to identify a function, and examples from the *otherwise* corpus. [LHS] and [RHS] correspond to the clause that is left and right of *otherwise*.

CONSQ.	ARG.	ENUM.	EXCPT.
.19	.45	.13	.26

Table 2: Function Distribution of the *Otherwise* corpus.

3.3 Function Signals

Our *otherwise* corpus contains naturally-occurring passages that are useful for corpus study. Particularly, we are interested in finding the signals that make a function available. We calculate pointwise mutual information (Torabi Asr and Demberg, 2013) for each word token w and discourse function l,

$$pmi(w, l) = \log \frac{p(w, l)}{p(w)p(l)}$$

A high PMI score indicates that word token w is highly associated with discourse function l, making the token a strong candidate for a lexical signal for that function. We only consider word tokens that occur in more than 15 passages to avoid overfitting the contents of the corpus (Zeldes and Liu, 2020).

Our results show that **modals** make up the largest group of signals. The functions they co-occur with depend on the modal type and its position: **Priority modals** (e.g., *need*, *must*, *should*) indicate how important and desirable an event is by the speaker (Pyatkin et al., 2021) and often occur in [LHS] to signal ARGUMENTATION:

> (1) Consumers should be told the whole truth. Otherwise, it amounts to fraud (CONSEQUENCE).

Plausibility modals (e.g., *could*, *can*, *may*), on the other hand, indicate how likely an event will happen given assumptions in the context (Pyatkin et al., 2021). Their appearance in [LHS] often indicates viability of an option and signals ENUMERATION, while *might, would, may* that appear in [RHS] often indicate the likelihood of an outcome and signal CONSEQUENCE or EXCEPTION:

> (2) The public can visit an exhibition to share their feedback. Otherwise, the public can submit feedback forms on the website. (ENUMERATION)
>
> (3) It's a good thing to ride horses at home and not at the racecourse. Otherwise, you might have been much more badly hurt. (CONSEQUENCE)

Other function signals include **negation markers** and **downward-entailing predicates**[2] in either [LHS] or [RHS] that indicate CONSEQUENCE and ARGUMENTATION:

> (4) She was nervous. Otherwise, she wouldn't be rambling. (CONSEQUENCE)
>
> (5) You keep your mouth shut, you never contact Nasry again, you don't lawyer up, this affidavit stays in a vault, and the video disappears. Otherwise, the charge will be murder. (ARGUMENTATION)
>
> (6) Generally, eating problems can be avoided by being flexible with your puppy from the start, varying the eating location, alternating types of dog food, and changing feeding times. Otherwise, be prepared for your dog to become accidentally conditioned by circumstances that lend new significance to the sound of the dinner bell. (ARGUMENTATION)

Focus particle *only* in either [LHS] or [RHS] that indicates ARGUMENTATION and EXCEPTION:

[2]Downward entailing constructions support valid reasoning from a set to a member. For example, John doesn't own a dog to John doesn't own a beagle (Webber, 2013).

(7) He spent only two years at school. Otherwise, he was educated at home. (EXCEPTION)

Connectives that appear at in [LHS] and not attached to *otherwise*[3]: *or* for ENUMERATION and *but* for CONSEQUENCE.

(8) Treatment may also be available in a Young Chronic Sick Unit or in a Geriatric Unit in a hospital. Otherwise, the patient might spend some time in a private nursing home. (ENUMERATION)

(9) Clearly, he resented Gavin but also had empathy towards him. Otherwise, he wouldn't have lent him money in the first place. (CONSEQUENCE)

We also found that several of these signals appear in function-bearing passages beyond those with sentence-initial *otherwise*. We gathered a substantially larger sample of passages ($n = 2656$) that contain *because otherwise*, *alternatively*, and phrases with *exception*[4], each marking distinct discourse functions—namely, CONSEQUENCE, ENUMERATION, and EXCEPTION. Across these passages, modal triggers, negation markers, and the connective *or* **remain** function signals.

While our data-driven method extracts words that co-occur with some *otherwise* functions, the method falls short in identifying discourse signals in the context that surrounds them, and establishing whether a comprehender might actually use them when inferring a function. To address this, we analyze the linguistic characteristics of tokens that a language model identifies as distinctive of a function (§4.3). As we will show, the model is sensitive to the context of an *otherwise* passage. In inferring the appropriate function, the model confirms the utility of several lexical signals we have identified in this section.

3.4 Metric: Continuation Acceptability

To study the capability of language models to distinguish *otherwise* functions, we propose a variant of surprisal-based metric, motivated by previous work on the semantics of generics and quantifiers (Cilleruelo et al., 2025). *Continuation acceptability* selects a discourse marker that indicates a distinct function and makes a continuation, [RHS], more likely given prior context, [LHS].

[3]In our corpus, we only consider sentence-initial bare *otherwise* without additional connective attached to it (e.g. *or otherwise*).

[4]Phrases to mark EXCEPTION are: *with the exception that*, *except for the fact that*, *an exception is that*, *one exception is that*, *as an exception*.

Definition Let D be a set of candidate discourse markers that are distinctive of a discourse function, (a_1, a_2), the [LHS] and [RHS] clause (or continuation) of a passage s in our *otherwise* corpus (with sentence-initial *otherwise* removed). We construct $\{(a_1 + d + a_2)|d \in D\}$, the set of variations of s, where $(a_1 + d + a_2)$ denotes string concatenation. d is the most acceptable discourse marker if its variation $(a_1 + d + a_2)$ has the lowest surprisal of continuation in a language model θ:

$$\arg\min_{d \in D} I[(a_1 + d + a_2); \theta]$$

where I is the surprisal of continuation:

$$I(s; \theta) = -\frac{1}{|A|} \sum_{i \in A} \log p_\theta(t_i | t_{<i})$$

with A the set of indices of tokens in continuation, t_i the tokens in passage s.

For example, consider the passage s, *Take your umbrella. Otherwise, you'll get wet.* Suppose we have candidate discourse markers *because otherwise* for CONSEQUENCE and *alternatively* for ENUMERATION. Then the respective variations for s are:

(1) Take your umbrella. *Because otherwise*, you'll get wet.

(2) Take your umbrella. *Alternatively*, you'll get wet.

The continuation acceptability for the first variant is calculated as the surprisal of [RHS], *you'll get wet*, conditioned on [LHS], *Take your umbrella*, and the candidate marker, *Because otherwise*. We expect the model to assign lower surprisal scores for the continuation when it is conditioned on the candidate marker of the correct function. Therefore, the model should assign lower surprisal to variation (1) than (2).

Notice that in all variations constructed, both the prior context and the continuation are kept the same. The only change is the candidate marker, which allows us to test its facilitating effect on the expectation for the continuation. Per-token surprisal also allows us to examine how language models respond to specific aspects of the continuation (§4.2, §4.3), which are more difficult to capture through mask-filling or prompting.

We use the average of per-token surprisal as indicated by the formula, but we also consider other aggregate functions for per-token surprisal in §4.2.

Function	Discourse Markers			
CONSEQUENCE	because if not	because [PRON] ¬[AUX]	because otherwise	
↳ ARGUMENTATION	unless this is done	(when/ by) failing to do so	for fear that	lest
ENUMERATION	alternatively	as an alternative	in addition	on the other hand
EXCEPTION	but mostly	but usually	but other than that	
CONTROL	otherwise			

Table 3: Candidate discourse markers and their corresponding function. [PRON] and ¬[AUX] correspond to the pronoun and negated auxiliary verb of [LHS].

4 Experiments

In Rohde et al. (2018b), human participants infer the discourse functions of an adverbial, which vary across passages. We raise the question of whether a language model can also discern varied interpretations of *otherwise* and accept annotators' assigned function. The experiments below use *continuation acceptability* (§3.4) to evaluate the capability of language models to do so. First, we validate that *continuation acceptability* captures the models' understanding of *otherwise* (§4.1). Then, exploiting per-token surprisal score, we explore what aspects of the continuation are important to identify the discourse function: we use alternative aggregate functions for per-token surprisal (§4.2) and linguistic annotations on tokens found to be distinctive of a function (§4.3).

4.1 Can continuation acceptability identify *otherwise* function?

Experimental setup. We used autoregressive language models of increasing size without further fine-tuning: GPT-2 Base (Radford et al., 2019), with 124 million parameters, and GPT-Neo (Black et al., 2021), with 1.3 billion parameters, and Mistral-7B-v0.1 (Jiang et al., 2023). We selected the GPT family because they are the standard models for testing psycholinguistic predictive power, allowing for comparability with prior work. We additionally included a newer open-weight model, Mistral-7B-v0.1, to access per-token surprisal.

We applied *continuation acceptability* (§3.4) to these models and our *otherwise* corpus.[5] Specifically, for a passage in the corpus and all candidate discourse markers, we calculate the *continuation acceptability* score of words in the continuation.

As shown in Table 3, we choose candidate discourse markers that are both relatively frequent and generally successful at capturing a unique *otherwise* function in naturally-occurring examples we sampled from COCA and BNC, which are also used for constructing the *otherwise* corpus. We select three or more candidate markers for each *otherwise* function to reduce the bias of syntactic constraints and specificity of one discourse marker (Pyatkin et al., 2023). We optionally allow *otherwise* to be a candidate and label its function CONTROL. When *otherwise* is chosen, the model doesn't prefer any marker that explicitly realizes a single function that we defined, but prefers the adverbial as is, in its original form.

For each candidate discourse marker [DM], we allow both inter-sentential *Arg1*. [DM] *Arg2* and intra-sentential concatenation *Arg1*(,) [DM] *Arg2* of a marker and optionally includes comma after [DM] if suitable. We choose the concatenation that is most accepted by the model for that marker.

	Researcher Label			
	Consq	Arg	Enum	Excpt
GPT2	.07	**.11**	.05	.04
GPT-Neo-1.3B	**.16**	.08	.05	.13
Mistral-7B-v0.1	.11	.11	.08	**.21**

Table 4: The proportion of passages predicting CONTROL (accept *otherwise* as Top 1 candidate discourse marker), corresponding to the researcher label in bold.

Results and Discussion. We consider the function to be correctly identified if a candidate marker of that function appears in the top $k = \{1, 3, 5\}$ accepted discourse markers. For example, a best-performing model will accept *but mostly*, *but usually*, and *but other than that* as top 3 markers for an EXCEPTION passage.

We first show that using surprisal to identify the *otherwise* function is not trivial. All models accept candidate discourse markers that explicitly realize a function more often than bare *otherwise* (i.e., without an additional conjunction before), which is the original discourse marker that appears in the passage (Table 4). The result suggests that these

[5] We use the discourse function assigned by the researcher as reference label.

$k =$		1	3	5
Majority		.45		
Mask Scoring	T5-Base	.45	.76	**.93**
Continuation Acceptability	GPT2	.51	.74	.85
	GPT-Neo-1.3B	.56	.78	.88
	Mistral-7B-v0.1	**.59**	**.80**	.89

Table 5: Overall passage accuracy using top $k = \{1, 3, 5\}$ predictions of discourse markers. **Majority** corresponds to assigning the majority function of the dataset to all passages.

	GPT2			**GPT-Neo-1.3B**			**Mistral-7B-v0.1**		
$k =$	1	3	5	1	3	5	1	3	5
Conseq	.36	.52	.61	.39	.64	.75	.41	.68	.84
Arg	.45	.75	.89	.44	.71	**.94**	.41	.73	.93
Enum	.41	.68	.81	.43	.78	**.86**	.59	.73	.84
Excpt	.40	.65	.77	.43	.64	.77	.48	.72	.79

Table 6: Per-function passage accuracy using top $k = \{1, 3, 5\}$ predictions of discourse markers.

models do not simply memorize the sentence-initial *otherwise*, and that each model favors a distinct function to be CONTROL except for ENUMERATION. All models accept bare *otherwise* less frequently for ENUMERATION than other functions. In other words, they accept candidate markers that explicitly indicate a function much more frequently (e.g., *alternatively* or *in addition*) than *otherwise*.

As there currently exists no system for identifying the discourse functions of *otherwise*, we use the majority function of the corpus as a baseline. In addition, we compare *continuation acceptability* against a mask-scoring baseline[6], defined as the model probability of inserting the connective d at the mask token between the arguments (a_1, a_2) (Kurfalı and Östling, 2021; Stodden et al., 2023):

$$P(d|a_1, a_2) \propto P_\theta(\texttt{[MASK]} = d|a_1\texttt{[MASK]}a_2)$$

We report accuracy where the function of any of the top-k predicted discourse markers matches the gold function. Table 5 and Table 6 show overall and per-function accuracy. Models using *continuation acceptability* outperform the majority baseline, and also surpass mask-scoring at top-1 prediction. Upon inspection, mask-scoring under-predicts EXCEPTION, achieving only 18.6% accuracy on EXCEPTION passages with top-1 predictions.[7] Moreover, in roughly 78% of cases, the model's top-1 prediction is bare *otherwise*, suggesting that T5-Base may have encountered and memorized these examples during training. However, mask-scoring achieves higher accuracy in identifying the gold function with top-5 predictions.

We also observe that a larger model doesn't guarantee better accuracy in prediction. Despite Mistral-7B-v0.1 having 7 times more parameters than GPT-Neo, its overall improvement is marginal. Additionally, Mistral seems to be biased toward interpreting *otherwise* as EXCEPTION at $k = 1$: we observe that Mistral infers the most EXCEPTION passages than other models (Table 7). It also most frequently infers bare *otherwise* in EXCEPTION passages across functions (Table 4).

Across all models, the most prevalent function of the Top 1 scoring marker corresponds to the researcher label (Table 7). Nevertheless, CONSEQUENCE passages are the most challenging of all functions, as all models predict CONSEQUENCE correctly less frequently at $k = 1$ compared to other functions. It is often confused with ARGUMENTATION, which is expected as ARGUMENTATION is a subordinate function of CONSEQUENCE, and they are semantically similar.

One concern is that a model may inherently prefer specific candidate markers, regardless of the passage, which would complicate our analysis of model competence in inferring a discourse function. We demonstrate that this is generally not the case in Appendix B.

Our results have shown that *continuation acceptability* can be used to identify *otherwise* function, though the success varies across discourse functions of the passage and models.

4.2 Are all tokens in continuation equally important to identify *otherwise* function?

Fang et al. (2025) have shown that for long context understanding, not all tokens are equally important to identify the answer token. Similarly, our corpus study (§3.3) finds that many lexical signals that help make functions available, such as modals and negation markers, often appear before the main verb. We hypothesize that not all tokens in a continuation are equally important for identifying the *otherwise* function and that mean surprisal

[6]We keep the same experimental setup but use T5-Base (Raffel et al., 2020), which is trained on a masked language modeling objective. We select this model because it supports multi-word predictions at the mask token, making it compatible with our candidate connectives. The top k candidate connectives with the highest probabilities are selected.

[7]When calculating per-function accuracy, we disregard predictions that are bare *otherwise*.

	Researcher Label															
	Consq				Arg				Enum				Excpt			
Top 1 Label	Consq	Arg	Enum	Excpt	Consq	Arg	Enum	Excpt	Consq	Arg	Enum	Excpt	Consq	Arg	Enum	Excpt
GPT2	**.36**	.32	.14	.18	.21	**.45**	.22	.12	.08	.35	**.41**	.16	.20	.23	.17	**.40**
GPT-Neo-1.3B	**.39**	.21	.27	.12	.29	**.44**	.17	.10	.16	.24	**.43**	.16	.12	.19	.27	**.43**
Mistral-7B-v0.1	**.41**	.30	.20	.09	.30	**.41**	.23	.07	.03	.27	**.59**	.11	.04	.27	.21	**.48**

Table 7: The distribution of Top 1 predicted label. A green cell indicates that the predicted function is acceptable, while a red cell indicates that it is unacceptable. We allow CONSEQUENCE candidate markers for ARGUMENTATION passages.

(i.e., token-level surprisal aggregate used in §3.4) may not be representative enough in predicting discourse function.

Experimental setup. In this experiment, we explore alternative aggregates for per-token surprisal and compare them with mean surprisal. We test both previously proposed aggregates in the literature (*superlinear*, *maximum*, *variance*, and *difference*; see Appendix C for full definitions) and new aggregates designed for our task. Specifically, we select key tokens using two criteria and average their per-token surprisal: 1) **Pre-root**: consider tokens up to the root (as defined in syntactic dependency) or head verb of the continuation and 2) **Most varied tokens (MVT)**: tokens with the largest variance in surprisal across variations of different discourse markers (that make a function available). We believe MVT allows us to pinpoint the exact location where the model diverges on its interpretation of *otherwise* function, given that the candidate marker is the only element that varies in our experiment. We consider the Top 3 most varied tokens. In addition to testing them separately, we also combine these two criteria.

We use GPT-Neo-1.3B for subsequent experiments, as it has comparable performance to Mistral-7B in our previous experiment and is often used in the psycholinguistic literature, which can shed some light on token-level understanding of the continuation.

Results and Discussion. Giulianelli et al. (2023) has shown that superlinear surprisal aggregate highly correlates with human acceptability judgments on an upcoming turn in dialogue from *Switchboard* (Godfrey et al., 1992) and *DailyDialog* (Li et al., 2017). Besides just dialogues, our results, which are tested on a wide range of genres (§3.2), confirm that superlinear is the best performing aggregate among what has been proposed in past literature, particularly in the context of continuation acceptability contingent on a function-indicating discourse marker.

Suprisal Metric		**Acc.**			**Avg # Token**
	$k=$	1	3	5	
Mean		.56	.78	.88	13.45
Superlinear ($n=0.5$)		.63	.83	.91	
Maximum		.51	.73	.85	
Variance		.52	.72	.85	
Difference		.54	.75	.87	
Ours					
Pre-ROOT	Top 3 MVT				
✓		.61	.81	.91	3.54
	✓	.54	.75	.88	3
✓	✓	**.64**	**.84**	**.93**	2.73

Table 8: Passage Top k accuracy where $k = \{1, 3, 5\}$ predictions of discourse markers with GPT-Neo-1.3B, using various per-token surprisal aggregates

Additionally, our proposed aggregates, **Pre-ROOT** or **Top 3 MVT**, obtain comparable or better performance when compared to mean surprisal, despite considering fewer tokens (around 3 tokens on average as opposed to $\approx$ 13 tokens). Particularly, when combining **Pre-Root** and **Top 3 MVT** criteria, the performance exceeds that of superlinear (Table 8). We also see that **Top 3 MVT** criteria itself doesn't filter tokens in a way to better identify *otherwise* function compared to mean surprisal. Upon examining the relationship between two criteria, we have found that on average 48% of the **Top 3 MVT** appear pre-root, and more so for CONSEQUENCE and ARGUMENTATION (51% and 49%): two functions that are associated with most types of lexical signals. Thus, we hypothesize that combining criteria **Pre-Root + MVT** provides linguistic cues to identify lexical signals that are predictive of a function. In what follows, we test this hypothesis by investigating linguistic information of

tokens that fulfill these two criteria and compare their characteristics with those of function signals found in our earlier corpus study (§3.3).

4.3 What lexical signals are predictive of *otherwise* function?

We would like to assume that the key tokens selected by our criteria are in fact relevant to the model's decision-making in predicting the *otherwise* function. In this experiment, we analyze the linguistic characteristics of tokens that the model identifies as distinctive of a function. We observe that some of these tokens confirm lexical signals we previously identified in the corpus study (§3.3), and they provide promising evidence on how model behavior, such as surprisal, can be useful for studying discourse signals.

Experimental setup. For each passage, we extract the following linguistic annotation for each token that is both **Pre-Root** and **Top MVT** (§4.2) [8]: 1) word type, 2) part of speech, and 3) dependency tag. For each type of linguistic annotation i, we calculate PMI score $pmi(i, l)$ as in §3.3, but extend i from word type to other linguistic information. For example, given the token *looking*, we calculate a score for its word type *look*, part of speech tag *gerund or present participle*, and dependency tag *root*.

A high PMI score indicates that the linguistic information i is highly associated with discourse function l as seen by the model.

Results and Discussion. We find that both word types of **modal tokens** and part-of-speech tag *modal* are high-scoring signals. PLAUSIBILITY modals (as defined in §3.3) *may, might, will* as a word type signal both CONSEQUENCE and ARGUMENTATION, while *can* and *could* signal ARGUMENTATION and ENUMERATION respectively. PRIORITY modals *need* and *must* signal ENUMERATION and EXCEPTION respectively. Interestingly, *modal* as a part of speech tag is only high-scoring for CONSEQUENCE and ARGUMENTATION.

We have found some other lexical signals that confirm those from the corpus study: **Negation** as a dependency tag signals consequence, while *no* and *nothing* as a word type signal EXCEPTION, and *not* signals CONSEQUENCE. **Focus particle** *only* as a word type signals EXCEPTION.

[8] with using spaCy `en_core_web_sm` pipeline, see https://spacy.io/usage/processing-pipelines

We also found lexical signals that were not previously discovered in the corpus study. For instance, the word type *become* is found to signal ARGUMENTATION, and there are eight of such instances where *become* occurs in `[RHS]` to express a new state when the situation in `[LHS]` doesn't arise:

> (1) It was essential that people try to connect. Otherwise, we would become a society of strangers. (ARGUMENTATION)

Because the language model we have chosen is auto-regressive (i.e., generates a continuation that is conditioned on previous context), we are unable to apply the same analysis on tokens in `[LHS]`. Nevertheless, it is reassuring to see that some key tokens extracted by the model confirm lexical signals we found in the corpus study, especially given the model likely has been exposed to far more *otherwise* passages than our corpus. More importantly, we show that the model has learned frequency-correlated cues during pre-training and assigns more varied surprisal on these tokens across candidate discourse markers that license a function. These findings provide some promising evidence that token-level surprisal may offer helpful information for future pragmatic studies. As a next step, stronger evidence for function signals could be obtained by directly manipulating them in the passage (e.g., ablation) while preserving the passage's meaning.

5 Conclusion

In this paper, we study the discourse functions of *otherwise* through language models. To do so, we introduce a new dataset (the *otherwise* corpus) and metric (*continuation acceptability*). With these tools, we show that language models exhibit some capability of inferring *otherwise* function, though their extent to do so varies across functions of the passage and the model. Additionally, we identify the types of lexical signals that influence the availability of specific discourse functions and reveal that the model attends to some of these signals when inferring the discourse function. We hope our findings open new doors for study on adverbial and discourse coherence in both psycholinguistic and computational research, and inspire developing pragmatically competent models.

Limitations

We acknowledge that our study has some limitations. First, our dataset only considers sentence-

initial *otherwise*. This helps us ensure the adverbial serves a discourse function and simplifies our data collection process. We recognize that this may not represent all use cases of the adverbial. It may also affect syntactic patterns and lexical signals of passages we have analyzed. For future research, we plan to collect passages where *otherwise* within a sentence serves a discourse function.

Second, our analysis was based on the assumption that surprisal scores from language models reflect human behavioral patterns such as reading time. Recent work has shown that as the model size of language models increases, when using surprisal, their psycholinguistic predictive power decreases. This may be because these models are exposed to much more data than humans are. We have chosen models that are highly correlated with human reading time in past studies (Cong et al., 2023) or are of moderate size. Nevertheless, more direct evidence for discourse coherence and surprisal could be obtained by collecting reading time data (with an emphasis on function signals and preverbal tokens) or calibrating large-size models with temperature-scaling (Liu et al., 2024), so that they are more predictive of human behavioral patterns.

Lastly, although there is clearly value in studying discourse functions of adverbials in the interest of discourse parsing and other natural language understanding systems, we have not pursued other potential roles of discourse function inferences. An extended study may examine the influence of adverbials and their discourse functions on other semantic and pragmatic phenomena such as conditionals, anaphora resolutions, and presupposition, all of which we believe to be relevant to *otherwise*.

References

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

BNC Consortium. 2007. The British National Corpus, XML Edition. Oxford Text Archive.

Gustavo Cilleruelo, Emily Allaway, Barry Haddow, and Alexandra Birch. 2025. Generics are puzzling. can language models find the missing piece? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6571–6588, Abu Dhabi, UAE. Association for Computational Linguistics.

Yan Cong, Emmanuele Chersoni, Yu-Yin Hsu, and Philippe Blache. 2023. Investigating the Effect of Discourse Connectives on Transformer Surprisal: Language Models Understand Connectives, Even So They Are Surprised. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 222–232, Singapore. Association for Computational Linguistics.

Mark Davies. 2008. The Corpus of Contemporary American English (COCA).

Yunfang Dong, Xixian Liao, and Bonnie Webber. 2024. Syntactic Preposing and Discourse Relations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2790–2802, St. Julian's, Malta. Association for Computational Linguistics.

Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. 2025. What is Wrong with Perplexity for Long-context Language Modeling? ArXiv:2410.23771 [cs].

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023. Information Value: Measuring Utterance Predictability as Distance from Plausible Alternatives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5633–5653, Singapore. Association for Computational Linguistics.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Oren Kalinsky, Guy Kushilevitz, Alexander Libov, and Yoav Goldberg. 2023. Simple and Effective Multi-Token Completion from Masked Language Models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2356–2369, Dubrovnik, Croatia. Association for Computational Linguistics.

Andrew Kehler. 2006. Discourse coherence. *The handbook of pragmatics*, pages 241–265.

Alistair Knott. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, University of Edinburgh.

Murathan Kurfalı and Robert Östling. 2021. Let's be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*,

pages 1–10, Online. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Tong Liu, Iza Škrjanec, and Vera Demberg. 2024. Temperature-scaling surprisal estimates improve fit to human reading times – but does it do so for the "right reasons"? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9598–9619, Bangkok, Thailand. Association for Computational Linguistics.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of What Art? A Call for Multi-Prompt LLM Evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949. Place: Cambridge, MA Publisher: MIT Press.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 367–379, Online. Association for Computational Linguistics.

Valentina Pyatkin, Shoval Sadde, Aynat Rubinstein, Paul Portner, and Reut Tsarfaty. 2021. The Possible, the Plausible, and the Desirable: Event-Based Modality Detection for Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 953–965, Online. Association for Computational Linguistics.

Valentina Pyatkin, Frances Yung, Merel C. J. Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. 2023. Design Choices for Crowdsourcing Implicit Discourse Relations: Revealing the Biases Introduced by Task Design. ArXiv:2304.00815 [cs].

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(140):1–67.

Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher N. L. Clark, Annie Louis, and Bonnie Webber. 2016. Filling in the Blanks in Understanding Discourse Adverbials: Consistency, Conflict, and Context-Dependence in a Crowdsourced Elicitation Task. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 49–58, Berlin, Germany. Association for Computational Linguistics.

Hannah Rohde, Anna Dickinson, Nathan Schneider, Annie Louis, and Bonnie Webber. 2018a. Exploring substitutability through discourse adverbials and multiple judgments. In *IWCS 2017-12th International Conference on Computational Semantics*. ACL Anthology.

Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018b. Discourse Coherence: Concurrent Explicit and Implicit Relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267, Melbourne, Australia. Association for Computational Linguistics.

Emily Sadlier-Brown, Millie Lou, Miikka Silfverberg, and Carla Kam. 2024. How Useful is Context, Actually? Comparing LLMs and Humans on Discourse Marker Prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 231–241, Bangkok, Thailand. Association for Computational Linguistics.

Abel Salinas and Fred Morstatter. 2024. The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4629–4651, Bangkok, Thailand. Association for Computational Linguistics.

William Berkeley Sheffield, Kanishka Misra, Valentina Pyatkin, Ashwini Deo, Kyle Mahowald, and Junyi Jessy Li. Is it JUST semantics? a case study of discourse particle understanding in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21704–21715. Association for Computational Linguistics.

Regina Stodden, Laura Kallmeyer, Lea Kawaletz, and Heidrun Dorgeloh. 2023. Using masked language model probabilities of connectives for stance detection in english discourse. In *Proceedings of the 10th workshop on argument mining*, pages 11–18.

Fatemeh Torabi Asr and Vera Demberg. 2013. On the Information Conveyed by Discourse Markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 84–93, Sofia, Bulgaria. Association for Computational Linguistics.

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. 2024. Surprise! Uniform Information Density Isn't the Whole Story: Predicting Surprisal Contours in Long-form Discourse. ArXiv:2410.16062 [cs].

Bonnie Webber. 2013. What excludes an Alternative in Coherence Relations? In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 276–287, Potsdam, Germany. Association for Computational Linguistics.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. Discourse Relations: A Structural and Presuppositional Account Using Lexicalised TAG. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 41–48, College Park, Maryland, USA. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN Language Models Learn about Filler–Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024. Prompting implicit discourse relation annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165, St. Julians, Malta. Association for Computational Linguistics.

Amir Zeldes and Yang Liu. 2020. A Neural Approach to Discourse Relation Signal Detection. *Dialogue & Discourse*, 11(2):1–33.

Zhi Min Zhou, Man Lan, Zheng-Yu Niu, Yu Xu, and Jian Su. 2010. The effects of discourse connectives prediction on implicit discourse relation recognition. In *Proceedings of the SIGDIAL 2010 Conference*, pages 139–146.

A Dataset Construction

Candidate passages When selecting the passages that contain a discourse adverbial, its occurrence is not sufficient. Sometimes, the adverbial does not function as a discourse marker, and instead modifies part of a syntactic matrix clause. Thus, we select the sentence that starts with the adverbial immediately after the previous sentence (i.e., sentence-initial *otherwise*). We find this strategy works quite well due to its syntactic convention. Our search patterns for COCA and BNC are `._Otherwise_,` where _ indicates a blank space [9]. In total, we have extracted 294 passages for function annotation from 7,770 passages in COCA and 1,014 passages in BNC.

Dataset annotation All candidate passages are annotated by the researcher. One-fifth of the passages is additionally annotated by four native or near-native adult English speakers.

For each candidate passage, the researcher prepares a paraphrase for each function shown in Table 1. The paraphrase selection is in two steps. A participant first selects one of three paraphrases for CONSEQUENCE, ENUMERATION, and EXCEPTION. If CONSEQUENCE is chosen, the participant is asked to accept or reject the ARGUMENTATION paraphrase to further distinguish this subordinate function. We report inter-annotator agreement between researcher and participant in Table 9.

The annotation is completed on the Qualtrics XM Platform.

	Consq/ Enum/Excpt	Arg Yes/No	All
Participant 1	0.79	0.83	0.83
2	0.89	0.87	0.89
3	0.95	0.80	0.91
4	0.84	0.84	0.86
Average	0.87	0.83	**0.87**

Table 9: Inter-annotator agreement (Cohen's Kappa) of *otherwise* functions implied by paraphrases between researcher and participant

We observe that the disagreements arise from participant bias toward an *otherwise* function or multiple interpretations of a passage. For example, of four instances where participants infer EXCEPTION and the researcher infers CONSEQUENCE, three instances come from Participant 1. The example below shows that multiple interpretations of a passage is possible:

> (1) But the problem was that I wasn't sure I could make it back to the hotel to catch my flight. Otherwise, I would have been game.

We believe this is one of the cases when both CONSEQUENCE and EXCEPTION may hold, as both paraphrases below are valid:

> (1a) But the problem was that I wasn't sure I could make it back to the hotel to catch my flight. **Because if I could have make it back in time**, I would have been game.
>
> (1b) **Generally**, I would have been game. **An exception is** that I wasn't sure I could make it back to the hotel to catch my flight.

B Candidate Discourse Markers and Their Continuation Acceptability Scores

For each candidate discourse marker, we provide the distribution of *continuation acceptability* scores from all models in Figure 2. There is no significant variation in the median, and this pattern is consistent across models.

C Surprisal Aggregates

Given a passage s in the order of main clause x, discourse marker d and *otherwise* clause y, a language model returns token-level surprisal for the continuation $s(y_t) = -\log p(y_t|y_{<t}, x, d)$. We then compare the predictive power of the following surprisal aggregates (Giulianelli et al., 2023) in inferring discourse functions:

Mean surprisal is the average of token-level surprisal over all tokens in y:

$$s_{\text{mean}}(y|x,d) = \frac{1}{N}\sum_{n=1}^{N} s(y_n)$$

Superlinear surprisal is the power sum of token-level surprisal, which indicates that a superlinear effect on y:

$$s_{\text{superlinear}}(y|x,d) = \sum_{n=1}^{N} [s(y_n)]^k$$

We experiment with $k = \{0.5, 0.75, ..., 5\}$

[9] COCA requires blank space between tokens.

Maximum surprisal is the maximum of token-level surprisal. It indicates that the most surprised token captures the overall surprisal of y:

$$s_{\text{max}}(y|x,d) = \max\left[s(y_n)\right]$$

Surprisal variance is the variance of token-level surprisal from the mean surprisal.

$$s_{\text{variance}}(y|x,d) = \frac{1}{N-1}\sum_{n=2}^{N}\left[s(y_n) - s_{\text{mean}}(y)\right]^2$$

Surprisal Difference is the sum of differences between contiguous token-level surprisal:

$$s_{\text{difference}}(y|x,d) = \sum_{n=2}^{N}\left|s(y_n) - s(y_{n-1})\right|$$

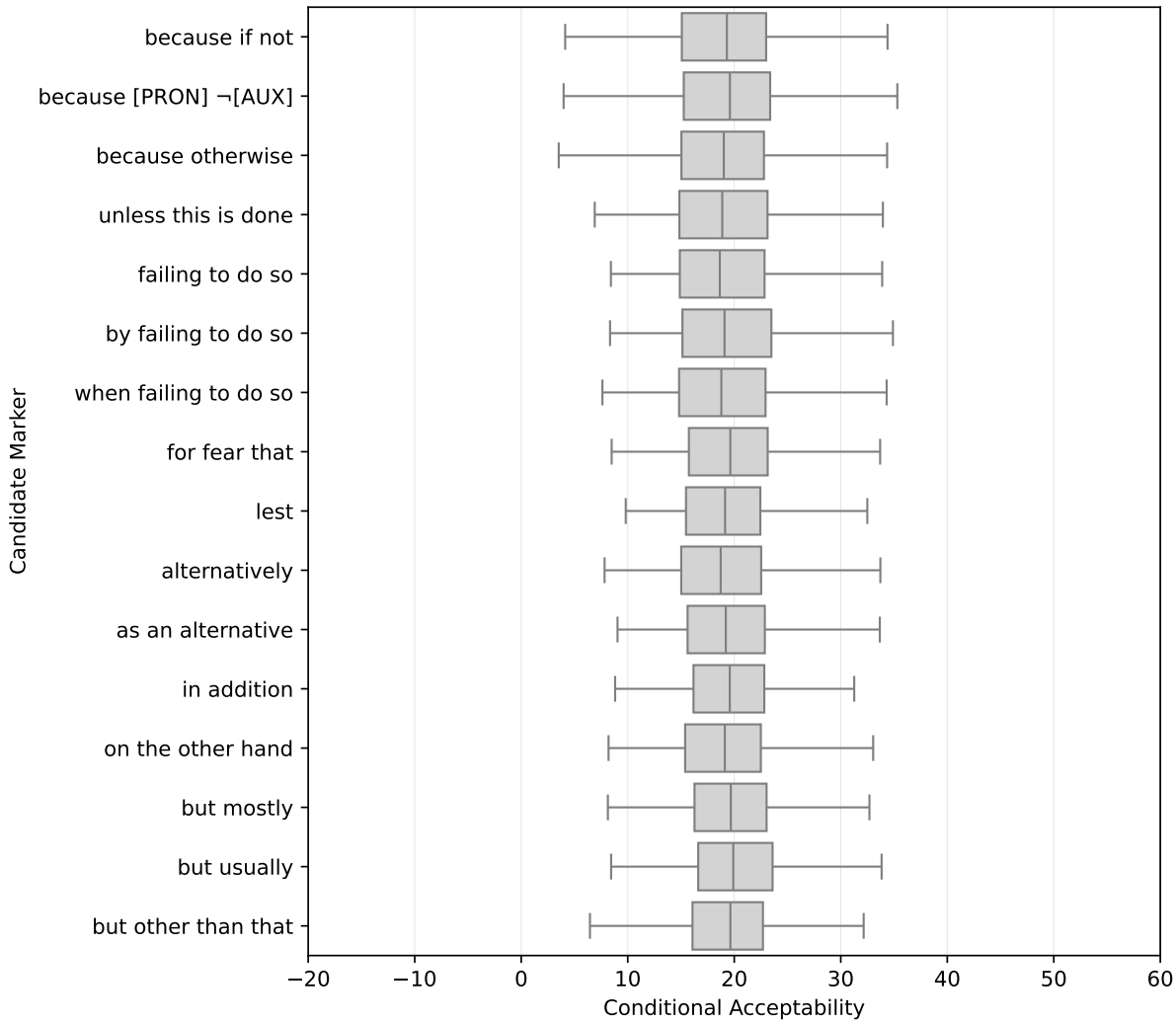

Figure 2: The distribution of *continuation acceptability* score of candidate markers