Causal Inferences Are Driven by Noun Concept Specificity: Evidence from Self-Paced Reading and Large Language Model Surprisal

Fabian Schlotterbeck*, **Raphael Barth***
University of Tübingen, Independent Researcher
fabian.schlotterbeck@uni-tuebingen.de, raphael.barthe@web.de
*co-first authors

Abstract

Inferring causal coherence relations has been argued to affect incremental language comprehension, particularly the integration of subsequent material. Building on prior psycholinguistic work, we investigated whether noun phrase (NP) specificity modulates the integration of explicit causal explanations. In a German self-paced reading study, we manipulated NP specificity in sentences containing implicit causality verbs, causally interpretable relative clauses, and explicitly causal subordinate clauses introduced by *because*. NP specificity systematically affected reading times in the *because*-clause, supporting the view that discourse-level causal expectations are derived incrementally. Within a frame-semantic account, more specific NPs activate more tightly constrained scenario structures, thereby strengthening causal expectations. Surprisal values from GPT-2 models replicated the critical interaction, though discrepancies with human reading times emerged. Overall, the findings support interactive, multi-component models of language processing and also speak to the relation between abstract psycholinguistic models and surprisal-based models of predictive processing affecting reading behavior.

Keywords: Causal Inferences, Coherence Relations, Elicitures, Noun Concept Specificity, Incremental Processing, Self-Paced Reading, Surprisal, Large Language Models

1. Introduction

Natural language involves drawing inferences at multiple levels. On the level of discourse comprehension one of the more elusive kinds of inferences are so called conversational elicitures (Cohen and Kehler, 2021). These inferences have been distinguished from entailments, presuppositions and implicatures. Two of their defining features are that they cannot be linked to a specific trigger and that they involve general cognitive biases, such as a tendency to draw causal connections between events. One classical case are causal inferences in discourse. In (1-a), for example, the relative clause tends to be understood as a causal explanation of the event described in the matrix clause although no unequivocal triggers for this inference can be identified. This is assumed to arise from the cognitive bias to construct causal relations between events, particularly when establishing coherence relations between discourse contents.

(1) a. The boss fired the employee who was always late for work.
b. The boss fired the employee because he was always late for work.

Relative clauses in this type of structure are often used restrictively to identify a referent. It is, however, assumed that the specific event described in the sentence in (1-a) in combination with the causality bias favor a causal coherence relation between the contents of the matrix and the relative clause. The resulting interpretation is similar to the interpretation of (1-b) where the relative pronoun *who* is replaced with *because he*.

Cases like these have not only attracted attention in theoretical linguistics but have also featured in psycholinguistic studies (Hoek et al., 2021; Rohde et al., 2011) where the interpretation of constructions like in (1-a) have been tested from the perspective of modularity (Fodor, 1983). In modular, as compared to interactive, processing models the processing of the phonological, morphological, syntactic and sematic features of phrases and sentences strictly precedes the assignment of coherence relations between them. In contrast to this hypothesis and in support of interactive architectures, empirical results discussed below showed that causal interpretation affects the incremental processing of the relative clause immediately.

Here we add to this line of work by taking a combined psycholinguistic and computational modeling perspective. Our contribution is twofold. Firstly, we identify another factor affecting the incremental processing of causal elicitures, besides the ones that have already been identified in previous research. In particular, we show that the specificity of the noun concepts that are involved affects incremental understanding of relative clauses as in (1-a), at least in German where verb information often follows rather than precedes the relative clause in comparable constructions. We discuss these results in relation to frame semantics (Fillmore, 1982) and predictive processing of coherence relations.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 276–285

Secondly, we show that critical effects related to the incremental assignment of coherence relations can be captured using surprisal calculated from the large language model (LLM) GPT-2, both in the German item set manipulating noun specificity and in an original item set manipulating the type of the relative clause, as well as its German translation. The second finding is partly inline with Surprisal Theory (Levy, 2008; Hale, 2001), which assumes that processing difficulty is determined by surprisal, but we also observe substantial deviation between surprisal and reading times. These findings are relevant regarding the questions how surprisal from LLMs is related to discourse processing architectures assumed in psycholinguistics and whether surprisal from LLMs suffices to capture sentence processing difficulty.

2. Previous studies on incremental processing of elicitures

In an ingenious design, Rohde et al. 2011 put interactive processing models to a tough test. They made use of known attachment preferences (formulated, e.g., in the Garden Path model of Frazier, 1987) of relative clauses in constructions such as (2) and tested whether these are modulated during online processing by causal expectations, induced, e.g, by certain types of verbs.

(2) a. John babysits the children of the musician who is generally arrogant and rude.
b. John babysits the children of the musician who are generally arrogant and rude.
c. John detests the children of the musician who is generally arrogant and rude.
d. John detests the children of the musician who are generally arrogant and rude.

According to the Garden Path model and supported by substantial empirical data, the relative clauses in (2) are preferably attached to the second, hierarchically lower NP, *the musician*, because this results in the simpler structure, as compared to high attachement to *the children*. The Garden Path model assumes modularity and, in particular, syntax-first processing (cf. Fodor, 1983; Chomsky, 1965) where initial syntactic parsing preferences are blind to other aspects of the sentence especially discourse-related effects. For these reasons, the plural disambiguation towards the dispreferred high attachment reading should cause processing difficulty. In contrast, to this hypothesis Rohde et al. 2011 adopted an interactive perspective (e.g. McRae et al., 1998; Tyler and Marslen-Wilson, 1977) and predicted that expectations for causal explanations, as induced, e.g., by the verb *detest*, can modulate parsing preferences incrementally. They contrasted implicit causality (IC) verbs like *detest*, which trigger strong discourse expectations for causal explanations involving the head NP, e.g. *the children*, with non-IC verbs like *babysit* that do not by themselves induce such explanations. The key idea is that the low-attachment preference is weakened to make room for a causal reading if an explanation is expected and high attachment can provide one. Rohde et al. 2011 thus predicted a smaller Garden Path effect on the number disambiguation in (2-c/d) than (2-a/b). This prediction was confirmed in a self-paced reading experiment providing evidence that causal coherence relations are inferred incrementally and may affect parsing preferences during online processing. This supports interactive processing models and challenges modular approaches like the Garden Path model.

Building on this design, Hoek et al. 2021 tested in an eyetracking during reading study whether causal inferences also modulate incremental processing of subsequent explicit causal relations as would be expected if they are drawn immediately, ruling out an explanation of the results of Rohde et al. 2011 based simply on the next-mention bias associated with IC-verbs (Bott and Solstad, 2023). They added a causal subordinate clause introduced by *because*, as in (3-a), in addition to the relative clause and examined whether causal expectations affect processing of the *because*-clause. In their study, causal expectations where modulated by the contents of the relative clause, which either functioned well as explanations as in (3-a/c) or not as in (3-b/d). If the relative clause is interpreted causally, integrating a second explanation introduced by *because* should be difficult because there arises a conflict between two potential causes in the discourse structure. They manipulated relative clause type (causal vs. neutral) and embedding connector (*because* vs. *so that*). While *because* introduces a causal explanation, *so that* introduces a consequential relation and should therefore not be sensitive to prior explanatory saturation.

(3) a. Diane fired the guy from the London office who was embezzling money because astoundingly he hired a stripper for the Christmas party.
b. Diane fired the guy from the London office who was here last month because astoundingly he hired a stripper for the Christmas party.
c. Diane fired the guy from the London office who was embezzling money and so astoundingly he hired a lawyer to sue the company.
d. Diane fired the guy from the London

office who was here last month and so astoundingly he hired a lawyer to sue the company.

The results revealed the critical interaction: with causal, but not consequential, subordinate clauses, causal relative clauses increased processing difficulty in the second subordinate clause. This effect was interpreted as evidence that an initial causal inference saturates the expectation for an explanation (in line with the Empty Slot Theory, Bott and Solstad, 2023), making a second explanation harder to integrate into the discourse model. Overall, these findings indicate that causal inferences, and elicitures more generally, are derived incrementally during online processing.

Hoek et al. 2021 found an early effect already on the adverb, e.g. *astoundingly*, directly following the connective. This effect was not particularly large in the measures directly related to first pass reading (they report $\beta = 59.40, SE = 29.52, t = 2.01, p < .05$ for regression path duration and a non-significant effect in first pass durations). The largest effect was found in total fixation durations and this could be driven by regressions initiated in the relative clause, which they did not analyze. That material within the subordinate clause is driving the effect is plausible since, when the connector is encountered, there are still ways to integrate another explanation (e.g. one that targets another level in the causal model). For example, the sentence in (4) provides a secondary explanation at another level without causing any conflict and this only becomes evident towards the end of the second subordinate clause. We will approach the question where in the sentence effects of causal inferences manifest in our surprisal-based simulations of the Hoek et al. 2021 data.

(4) Diane fired the guy from the London office who was embezzling money because obviously she hates bad money-management.

3. Previous work on LLM surprisal as a model of processing difficulty during reading

Surprisal theory (see Hale, 2001; Levy, 2008, for two different conceptions) equates processing difficulty with the contextual probability of an incoming linguistic unit, formally defined as the negative logarithm of its conditional probability, $-\log P(w_i|context)$. Lower-probability words incur higher processing costs, and difficulty is thought to be driven by to the reallocation of probability mass over competing structural analyses upon encountering new input. Empirically, surprisal correlates robustly and approximately linearly with reading times (e.g. Smith and Levy, 2013), capturing predictability effects and many aspects of syntactic ambiguity resolution (e.g. Arehalli et al., 2022).

Recent work using transformer-based language models (Vaswani et al., 2017) further refines this picture. While larger models typically achieve lower perplexity (i.e. better next-word prediction), their surprisal estimates may fit human processing measures worse, leaving more variance unexplained (Oh and Schuler, 2023). One explanation that has been proposed is that larger models utilize an unrealistic amount of knowledge in their next-word predictions which surpasses human capacities employed during real time processing. In fact, medium-sized models such as GPT-2 have been argued to occupy a sweet spot: they capture broad distributional regularities relevant to human expectations but not to an unrealistic degree, thus trading-off complexity and predictive accuracy. As a result, GPT-2–based surprisal seems to align more closely with human reading times than surprisal derived from substantially larger models. Human-like predictive behavior may depend not only on overall predictive accuracy but also on aspects like capacity constraints. For this reason we use GPT-2 in our simulations below.

At the same time, surprisal-based models sometimes underestimate the magnitude of processing disruption, especially in cases that have gained substantial attention in psycholinguistic studies. An important example are garden-path effects (e.g. Arehalli et al., 2022) where comprehenders are assumed to strongly commit to an initially preferred parse (e.g. the low attachment preferenmce discussed above). The reason may simply be that predicted difficulty at disambiguation points is attenuated because expectation-based models distribute probability mass across a larger set of analyses in parallel, relative to humans. Earlier commitments to an interpretation in humans vs. LLMs may be incentivized by the fact that, during conversation, language is not only parsed but also used in subsequent cognitive processes, involving, e.g., inference drawing or decision making. Another reason why processing disruption may sometimes be underestimated by surprisal estimates maybe that aspects related to memory-based costs (involved, e.g., in retrieval or reanalysis operations) contribute to the relatively large disambiguation penalties in humans (see, e.g., Oh and Schuler, 2022, for a hybrid approach incorporating such aspects).

Against this background, it is interesting to see whether the subtle coherence-related effects described in the previous section and also the data from our own self-paced reading study are captured well by LLM surprisal. This is for several reasons. Firstly, causal inferences may be part of the just-mentioned extra cognitive processes

that may not be reflected in surprisal and may encourage early processing commitments in humans. Causal elicitures are, moreover, thought to depend on general cognitive biases related to causal world and discourse models that need not be reflected in LLM surprisal. Finally, memory processes may be involved in coherence-related effects because at least two propositional contents need to be retrieved from or held in memory in order to assign a coherence relation. If this happens incrementally while a sentence is parsed, the memory burden may be especially large. Therefore, causal elicitures pose an interesting challenge to Surprisal Theory.

4. NP specificity and the design of the current study

In the present study, we carried out a self-paced reading experiment combining central ingredients from previous research with an additional crucial manipulation, namely the specificity of the NP concepts involved. The underlying idea is that more specific NP concepts foster stronger activation of scenarios linked to the causal expectations triggered by IC verbs (in the broadest sense). Verbs and their arguments are assumed to be associated with prototypical scenarios, called frames (Fillmore, 1982), and causal relations often arise within such frames. More specific NPs are linked to a smaller set of frames and thus causal expectations arising form within one of these frames may exert a larger effect as compared to less specific NPs, which are linked to a larger set of frames. Thus, just as verb type, relative clause type and connector type modulate effects related to causal inferences (see previous research section), the specificity of NP concepts may exert a comparable influence.

To test this, we adopted the items from Hoek et al. 2021 that contained *because*-clauses and added a manipulation of the NP specificity, as in (5-a/c) vs. (5-b/d) (all followed by (5-e)). The experiment was conducted in German, which resulted in translations with a different word order: the verb followed the NPs and the relative clause. Consequently, at the point of processing the relative clause, it is primarily the NP information that determines whether a causal interpretation is established. Following the verb, and parallel to the design of Hoek et al. 2021, a subordinate clause introduced by *weil* (because) was presented, and processing difficulty was measured in this clause. Across all experimental items, three regions of interest for self-paced reading were defined after the connector. These were: a pronoun, e.g. *er* (he), an adverbial, e.g. *plötzlich* (suddenly) and the rest of the clause.

(5) a. Die Vermieterin hat den Mieter, der
the landlady has the tenant who
letztes Jahr ihre Wohnung
last. year her apartment
überschwemmt hat, verklagt
flooded has sued
'The landlady sued the tenant who flooded her apartment last year'

b. Die Frau hat den Mann, der
the woman has the man who
letztes Jahr ihre Wohnung
last year her apartment
überschwemmt hat, verklagt
flooded has sued
'The woman sued the man who flooded her apartment last year '

c. Die Vermieterin hat den Mieter, der
the landlady has the tenant who
letztes Jahr eine Wohnung gesucht
last year a apartment sought
hat, verklagt
has sued
'The landlady sued the tenant who was looking for an apartment last year'

d. Die Frau hat den Mann, der
the woman has the man who
letztes Jahr eine Wohnung gesucht
last year a apartment sought
hat, verklagt
has sued
'The woman sued the man who was looking for an apartment last year'

e. ...weil er plötzlich anfing im
...because he suddenly started in_the
Gebäude zu rauchen.
building to smoke.
'...because he suddenly started smoking in the building.'

In addition to the self-paced reading study, we computed surprisal values from LLMs, in particular a German version of GPT-2, for the connector and the ROIs inside the *because*-clause. We conducted such a surprisal-based analysis of the English Hoek et al. 2021 items, their German translation and our German items with the NP specificity manipulations. This served three purposes. First, we asked whether the coherence-related effects under investigation are reflected in surprisal, which constitutes an interesting question in its own right (see discussion in the previous section). Second, we aimed to assess whether the effects observed in the original items from Hoek et al. 2021 would also emerge in our translated German materials, potentially reflecting the manipulation of NP specificity. Third, we sought to further localize the source of the effects by examining where within the relative clauses the

relevant differences arise. Since effects in the Hoek et al. 2021 study were mainly driven by total fixation duration, it is possible that effects in surprisal, which have been linked to first pass reading (Smith and Levy, 2013; Shain et al., 2024), only show up relatively late within the relative clause. This would be in line with the idea that integration is only possible and difficulty of the integration process is only determined after sufficient information about the content of the *because* clause has been processed, as in example (4). Following the the chronological order in the study, we report the surprisal-based analyses first and then present the reading time experiment.

5. Methods

5.1. Surprisal Simulation

Surprisal was computed using the *llm-surprisal*[1] tool with the pretrained models *gpt2-large*[2] (774M paramteres) and *german-gpt2-larger*[3] (137M parameters) with the default temperature setting. For each item, token-level surprisal values (negative log-probabilities) were extracted and summed within each predefined region of interest (ROI). Thus, the reported values reflect cumulative surprisal per region rather than mean per-token surprisal. Under a surprisal-based account of incremental processing, cumulative surprisal within an ROI is predicted to be proportional to processing difficulty.

5.2. Self-paced reading experiment

5.2.1. Participants and procedure

100 participants were recruited over prolific.co and redirected to `pavlovia.org`, where the experiment started automatically in full-screen mode. Participants were distributed randomly to one of four lists. They read sentences in a self-paced reading paradigm using the moving-window technique (Just and Carpenter, 1976). The latency between key presses served as the reading time measure. After the final word of a sentence, a comprehension task followed. Responses were given using the left arrow key ("yes") or the right arrow key ("no"). The next trial then commenced automatically. Participants gave informed consent.

The experiment began with a brief introduction and detailed instructions. Participants were explicitly informed that the comprehension questions following each sentence had a single unambiguous correct answer and that no ambiguity was intended. They were further advised that previously presented words could not be revisited. At the end of the experiment, participants were debriefed and provided with a completion link to confirm participation. The total duration of the experiment was approximately 15–20 minutes. Participant received 5 € compensation.

5.2.2. Materials

We constructed 28 items by translating the conditions with *because*-clauses from Hoek et al. 2021 into German. In doing so, several adjustments were necessary. First, we modified the structure of the embedded clause, as the succession of elements differs between English and German due to word order differences. We ensured that, following the connective, each item contains (i) a pronoun (e.g., *er*), (ii) an adverbial (e.g., *offensichtlich*), and (iii) a longer phrase, which constitutes the third region of interest (ROI). Second, the word order in the matrix clause differs from the original English examples. In our German materials, the structure is as follows: first NP, second NP, relative clause, and finally the verb. Third, we systematically adjusted the specificity of the noun phrases. Since specificity is not fully consistent in the original items, we established a constant level of specificity in our base versions (role nouns as in (5)). Based on these translations, we constructed an additional set of conditions in which the noun phrases are less specific (either generic nouns like *man* and *woman* in (5) or proper names). If the specificity manipulation could not be carried out in a satisfactory way in a given item, it was removed from the list and replaced by an item that was generated from scratch (7 in total). An example item is provided above in (5). Overall, this yielded a 2 × 2 factorial design (Relative Clause Type × NP Specificity) with both factors manipulated within items and within participants. The 28 experimental items were distributed over four lists according to an Latin square. They were accompanied in each list by 27 filler items. In addition, 32 items from a related experiment were added which contained also causal inferences but of a different kind, namely the causal interpretation of a adjectival modification (*The drunk/young driver had an accident*).

5.3. Statistical analyses

Surprisal values and reading times were both analyzed using linear mixed-effects models.[4] These

[1] https://github.com/tmalsburg/llm_surprisal

[2] https://huggingface.co/openai-community/gpt2-large

[3] https://huggingface.co/stefan-it/german-gpt2-larger

[4] Data and analysis scripts will be made available at https://osf.io/rckpt/overview?view_only=d99e4d5aedea40698790000efc52494f

models were fit using the lme4 package (Bates et al., 2015) for the statistical computing software R. Besides fixed effects for the experimental manipulations and their interaction, they included random intercepts for items and (if applicable) participants as well as random slopes for the experimental manipulations (without interaction) provided this allowed for convergence and did not lead to overfitting. We analyzed the four ROIs in the causal (*because*) or consequential (*and so*/*woraufhin*) clause (i.e. the connector, pronoun, adverb and rest) and for the simulation study we also analyzed the cumulative surprisal in all four ROIs if no effects in the individual ROIs were found. We computed one-tailed tests (with $\alpha = .05$), as we tested for a directed hypothesis, i.e. a specific type of interaction consisting in reduced processing disruption in *because*-clauses after conditions with restrictive relative clauses assumed to weaken causal inferences. In addition to the model specifications, all summaries of statistical models in all ROIs with significant effects and corresponding plots are provided in the accompanying repositories.

For the self-paced reading experiment, the data were cleaned in three steps. First, the entire distribution of reading times for each participant were inspected visually to spot unusual behavior. This led to the exclusion of one data set which seemed to be of non-human origin because it showed a multimodal RT distribution with five modes almost constantly across the experiment. Second, participants were removed that answered more than 5 out of 20 designated control questions in filler trials incorrectly. This affected not a single participant. Second, trials with RT shorter than 200 ms or more than 2.5 standard deviations above the mean in each ROI were removed. This affected 9.2% of the data.

6. Results

6.1. Surprisal simulation

Analyses of the surprisal values revealed the critical interaction in each of the three item sets. In the original item set from Hoek et al. 2021 (cf. example (3)), the critical interaction was found immediately on the connector *because* / *so that* ($t = -1.905, p = .033$). In that region, restrictive relative clauses led to higher surprisal than causal ones, but this difference was smaller for causal vs. consequential continuations (see Figure 1). The critical interaction was, furthermore, marginal in the last ROI containing all the words after the pronoun ($t = -1.487, p = .071$). In that region, we observed a cross-over pattern where causal relative clauses led to numerically higher surprisal in causal continuations than restrictive ones, but the opposite pattern was observed for consequential continuations.[5]

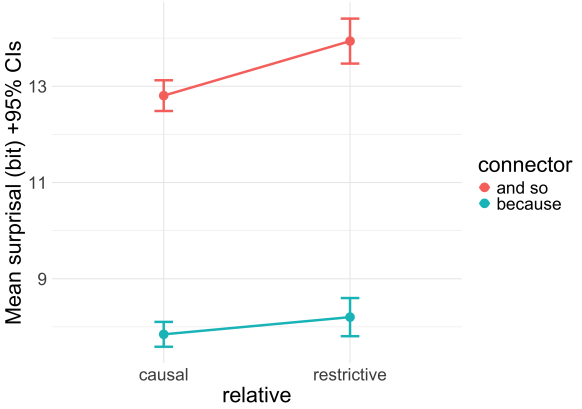

Figure 1: Surprisal in the connector ROI of the Hoek et al. 2021 items.

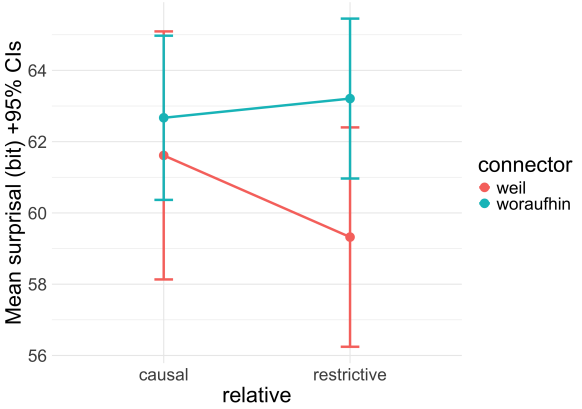

Figure 2: Cumulative surprisal over the entire subordinate continuation clause in the German translation of the Hoek et al. 2021 items. *weil*: 'because'; *woraufhin*: 'and so'.

In the German translation of the item set from Hoek et al. 2021, the critical interaction was not significant in any of the individual ROIs. However, when the connector and the three following regions were analyzed together and cumulative surprisal was computed across all four regions, the predicted interaction emerged ($t = -1.89, p = .034$; see Figure 2). The interaction was due to a crossover pattern where causal relative clauses led to lower surprisal in consequential continuations but higher values in causal ones.

The simulation of the NP Specificity manipulation

[5] A first surprisal analysis of these items was presented in the M.A.-thesis of the second author (Barth, 2025). This analysis was based on a smaller version of gpt-2[6] (with 124M parameters) and it only revealed the critical interaction in cumulative surprisal values over the entire subordinate clauses. In that analysis causal relative clauses led to numerically lower surprisal in consequential continuations but higher values in causal ones.

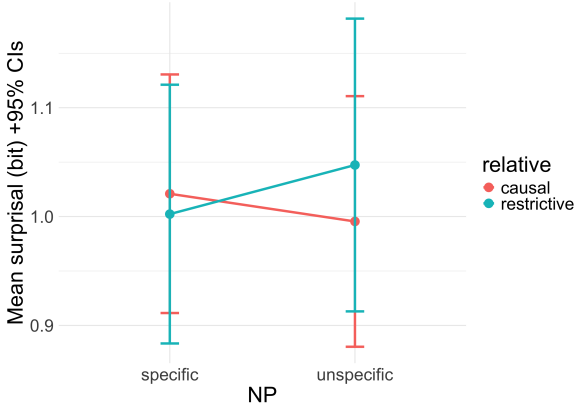

Figure 3: Surprisal in the ROI containing the pronoun for the NP-specificity effect in German.

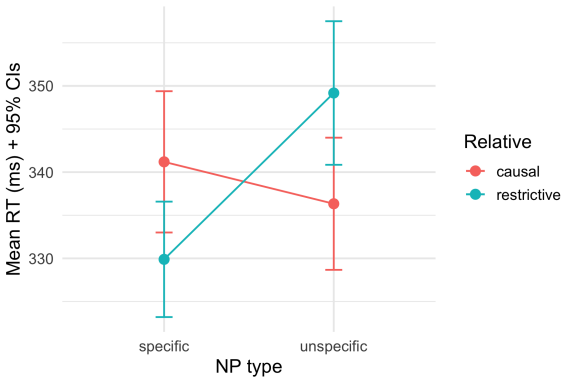

Figure 4: The NP specificity effect in German. Mean reading times and 95% CIs for the adverbial region in the *because* clause.

revealed the critical interaction in the ROI containing the pronoun (see Fig 3; $t = 2.603, p = .007$). In this region, specific NPs led to higher surprisal in combination with causal as compared to restrictive relative clauses whereas the opposite pattern was observed for unspecific NPs. The same numerical pattern was already present on the connector (*because*) preceding the pronoun, but it did not lead to a significant interaction effect in that ROI ($t = 1.583, p = .063$).

6.2. Self-paced reading

Significant effects in the reading time analysis were restricted on the second spillover region, i.e., the adverbial, for example, *plötzlich* ('suddenly'). In this region, the predicted interaction between NP Specificity and the Relative Clause Type, restrictive vs. causal, was significant ($t = 1.702, p = 0.045$). For the unspecific NPs, restrictive relative clauses led to longer reading times than causal ones, whereas the opposite was observed for the specific NPs (see Figure 4). No other effects were significant (all $|t| < 1$).

7. Discussion

In the present study, we built on previous research on causal inferences (Rohde et al., 2011; Hoek et al., 2021), which have been termed conversational elicitures in theoretical linguistics (Cohen and Kehler, 2021) and identified as a central type of inference in natural language. Psycholinguistic evidence has shown that such causal inferences are derived incrementally during online processing and affect the interpretation of upcoming material. In particular, once a causal explanation has been inferred, additional explicit explanations introduced by *because*-clauses become more difficult to integrate, as reflected in increased reading times. Extending these findings, we manipulated the specificity of the noun concepts involved. We observed in our self-paced reading study that NP specificity modulated the difficulty of integrating an upcoming *because*-clause.

We interpret this effect within a frame-semantic perspective (Fillmore, 1982). On this view, verbs and also the noun phrases filling their thematic roles are associated with prototypical scenarios, or frames. More specific noun phrases (e.g., *landlord*) evoke a narrower and more strongly constrained set of such scenarios than less specific noun phrases (e.g., *woman*). Since causal inferences crucially rely on these scenarios as an interface between linguistic and world knowledge, stronger scenario activation should amplify causal expectations. Accordingly, when a specific NP strongly activates a limited set of frames, additional causal material may lead to increased integration costs due to expectation saturation.

Our results show that NP specificity modulates the difficulty of integrating further causal explanations in a systematic way. What was, however, unexpected in the reading-time data was the observation that, in combination with unspecific noun phrases, the causal relative clause lead to faster integration of the *because*-clause than did the restrictive relative clause. One explanation for this observation could be that the cue provided by unspecific noun phrases may be too weak to trigger a strong causal inference. Given the general bias toward establishing causal coherence relations, the explicit *because*-clause may therefore enhance discourse felicity and be processed relatively easily, as it satisfies an independently strong expectation for causal coherence.

Turning to the surprisal-based modeling, we would like to highlight and discuss two central observations. First, across all simulations we conducted, i.e. the original dataset from Hoek et al. 2021, the German translations of these items, and our newly

constructed materials manipulating NP specificity, the critical interaction that has been taken as evidence for causal inferences affecting incremental processing was found. In this respect, LLM-based surprisal provides a useful model of reading behavior in a theoretically important domain. Causal inferences are particularly relevant here because, unlike phenomena like core syntactic parsing, semantic processing or certain pragmatic inferences, they operate at the interface between linguistic and world knowledge. The fact that LLM surprisal captures the relevant interaction suggests that predictive language models are sensitive to discourse-level regularities of this kind.

Second, despite the observed convergence, systematic discrepancies between human reading times and surprisal-based estimates were observed. In particular, there was a cross-over pattern in the reading time effects reported by (Hoek et al., 2021) that was absent from our surprisal estimates. In their data, the conditions with causal continuations led to a slow down when they contained causal as compared to restrictive relative clauses whereas the opposite effect was observed in conditions with consequential continuations. In addition, the critical effects were observed one region earlier in our suprisal estimates than in the reading time data sets. This is not particularly surprising given that surprisal has been shown empirically to affect reading behavior in spillover regions (e.g. Smith and Levy, 2013). One possible explanation for discrepancies in the pattern of effects could thus be that effects on reading behavior are determined by the surprisal of the previous word as well as processes directly related to the current region. In addition, the phenomena under investigation may recruit cognitive resources that go beyond predictive processing. In particular, memory-related processes such as maintaining previously processed sentence material and linking it to scenarios retrieved from long-term memory may contribute to integration costs in ways that are not directly reflected in surprisal (e.g. Oh and Schuler, 2022). Moreover, human reading behavior may also be shaped by adaptation effects arising within an experimental setting. Participants are repeatedly exposed to structurally and semantically similar sentences, which may lead to adaptation over the course of the experiment, as shown in previous research (Fine et al., 2010). By contrast, surprisal is computed independently for each item, without any possibility of adaptation.

Moreover, our surprisal estimates suggest that some of the effects in the study of Hoek et al. 2021 may not be strictly local. While they reported relatively early effects, the largest effects they found were in total fixation duration. Surprisal estimates provided some indication that part of this effect may be driven by surprisal values of the final region of the sentence. This is also corroborated by the results in our German translation of their items, where no reliable effects emerged in the individual ROIs but the critical interaction only emerged when surprisal was summed across the entire subordinate continuation clauses. In this case, the system appears to accumulate information across the clause before integration difficulty becomes fully manifest. From a theoretical perspective, this gradual accumulation is plausible, as the incompatibility between competing cues may only become evident once sufficient propositional content has been processed (cf. the discussion of example (4) above).

Our data also provide some information regarding discussions about the type of LLM-surprisal that is best suited for capturing reading time effects. In analyses based on reading time corpora, Oh and Schuler 2023 found that smaller models, e.g. smaller GPT-2 models, provide the best fit for reading time data. In our study, we used smaller models for the German than for the English items. In the second simulation, i.e. in the simulation of our German translation of the items from Hoek et al. 2021, the smaller models only revealed the critical interaction in cumulative surprisal across the entire subordinate continuation clause. The same general pattern was also found in previous analyses conducted in the context of the M.A. thesis of the second author (Barth, 2025), who used a similar-sized model for the original English items and also found the critical interaction only in cumulative surprisal values (cf. the analyses provided in the accompanying repository). In contrast, the larger GPT-2 model revealed early effects already on the connector, in line with the study of Hoek et al. 2021. One possibility is that with regard to the type of discourse-related effects we studied here an increase in model size from the smaller GPT-2 variants may yield a closer match to reading time data.

Besides model size, another dimension that has been discussed in the literature is the layer from which surprisal is extracted. All our analyses as well as the systematic comparison by Oh and Schuler 2023 were based on surprisal from the final layer of LLMs. In a recent study, Kuribayashi et al. 2025 showed, however, that surprisal extracted from internal layers of larger models captures data from reading time corpora as well as surprisal from the final layer of smaller variants or even better.[7] An interesting question for future research would be whether this also holds for the type of discourse-related effects studied here.

With regard to the debate surrounding modular vs. interactive psycholinguistic processing models, our data are somewhat ambiguous. While they

[7] We would like to thank an anonymous reviewer for pointing this out to us.

show that effects related to discourse coherence affect incremental language processing, in line with interactive models, we also found some indication of effects that emerge relatively late in the relevant *because*-clauses, as would be expected from a modular perspective. We would like to propose though that our data are more compatible with a nuanced interactive perspective where effects due to the integration of causal relations into the discourse model show up immediately but are still modulated by information later on, reflecting the vast space of potential unfolding discourse structures and resulting in no immediate effects already on the connective.

8. Acknowledgements

We would like to thank three anonymous reviewers for helpful comments on an earlier version of this paper. We also thank Irene Rapp, Stefan Engelberg, and Robin Hörnig for helpful comments and discussion on the M.A. project of Raphael Barth on which the paper is based. FS received funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 538184825.

9. Bibliographical References

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL 2022)*. Association for Computational Linguistics.

Raphael Barth. 2025. Psycholinguistische Untersuchung des Einflusses von Nominalphrasenkonzepten auf die kausale Inferenzbildung. Master's thesis, University of Tübingen, Tübingen, Germany.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Oliver Bott and Torgrim Solstad. 2023. The production of referring expressions is influenced by the likelihood of next mention. *Quarterly Journal of Experimental Psychology*, 76(10):2256–2284.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge.

Jonathan Cohen and Andrew Kehler. 2021. Conversational eliciture. *Philosophers Imprint*, 21(12):1–26.

Charles Fillmore. 1982. Frame semantics. *Linguistics in the Morning Calm. Seoul*, pages 111–137.

Alex Fine, Ting Qian, T. Florian Jaeger, and Robert Jacobs. 2010. Syntactic adaptation in language comprehension. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, pages 18–26, Uppsala, Sweden. Association for Computational Linguistics.

Jerry A Fodor. 1983. *The modularity of mind*. MIT press, Cambridge, MA.

Lyn Frazier. 1987. Sentence processing: A tutorial review. In Max Coltheart, editor, *Attention and Performance 12: The Psychology of Reading*, pages 559–586. Lawrence Erlbaum Associates.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American chapter of the Association for Computational Linguistics*.

Jet Hoek, Hannah Rohde, Jacqueline Evers-Vermeul, and Ted J.M. Sanders. 2021. Expectations from relative clauses: Real-time coherence updates in discourse processing. *Cognition*, 210:104581.

Marcel Adam Just and Patricia A Carpenter. 1976. Eye fixations and cognitive processes. *Cognitive Psychology*, 8(4):441–480.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *Transactions of the Association for Computational Linguistics*, 13:1743–1766.

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

Ken McRae, Michael J Spivey-Knowlton, and Michael K Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in online sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Hannah Rohde, Roger Levy, and Andrew Kehler. 2011. Anticipating explanations in relative clause processing. *Cognition*, 118(3):339–358.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Nathaniel J. Smith and R. Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–19.

Lorraine K. Tyler and William D. Marslen-Wilson. 1977. The on-line effects of semantic context on syntactic processing. *Journal of Verbal Learning and Verbal Behavior*, 16(6):683–692.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.