Abstract

The purpose of this work is to investigate the sensitivity of Large Language Models (LLMs) to garden-path sentences. The latter are structurally correct sentences which however elicit a *reanalysis* effect after the "parser" realizes the preferred analysis is not correct (Frazier, 1979; Frazier & Rayner, 1982).

In order to assess whether language models find garden-path sentences surprising, we measured the perplexities that GPT-2 assigned to a corpus of garden-path sentences and compared the scores with those assigned to their corresponding baselines. Furthermore, we investigated whether the model's responses varied significantly across different types of garden-path sentences.

The results are in line with our predictions: garden-path sentences are systematically more confusing to the model than their baseline counterparts. Moreover, among the four garden-path subtypes considered, sentences of the MV/RC type turned out to be the most perplexing with respect to the baselines, followed by Categorical ambiguity sentences, NP/Z sentences and NP/S sentences. These results are indicative of the model's capabilities in predicting garden-path effects observed in humans.

Contents

List of figures

List of tables

Investigating LLMs' predictions of garden-path effects

Introduction:

Large Language Models (LLMs) are often referred to as *black boxes*, due to the inaccessibility of their internal states. Researchers in NLP have devised a variety of methods for bypassing this limitation and indirectly probing their hidden states, with the goal of enhancing the interpretability of their outputs.

This work intends to continue this line of research by investigating the perplexity patterns of GPT-2 relative to *garden-path* sentences, and checking whether they align with the human processing difficulties traditionally associated with this linguistic phenomenon (e.g., Frazier, 1979; Frazier & Rayner, 1982). It aims to do so by means of a simple behavioral test. First, a dataset of garden-path – baseline sentence pairs was hand-crafted or collected from relevant works. Second, the model's responses to the presented stimuli were observed and perplexity scores calculated. Furthermore, the given garden-path sentences were also sorted into categories, in order to examine their effect on the model's perplexities.

Even if behavioral tests do not try to access a language model's internal representations as deeply as other methods (such as probing), they can still provide valuable information. Specifically, this experiment could offer insight into the model's ability of correctly predicting garden-path effects.

This work is articulated as follows:

- The first part of the paper presents the theoretical background within which this work is situated: on the one hand, a brief overview of the literature on interpretable NLP and on the treatment of garden-path effects in this field is provided (§ 1.1); on the other hand, the classical account of the garden-path phenomenon in psycholinguistics literature is summarized in § 1.2.
- The second part delves into the details of the experiment: § 2.1 outlines the methodological steps that were followed for building up the experiment, while § 2.2 discusses the experimental results in depth.

1. Theoretical background

This first part of the paper defines the theoretical framework that influenced this paper: in § 1.1, relevant literature in NLP is briefly reviewed, while in § 1.2 the psycholinguistic account of the garden-path phenomenon is presented.

1.1. Overview of interpretable NLP and current treatment of garden-path effects:

As already mentioned above, LLMs' inner workings are not accessible for direct examination. Therefore, researchers have developed different ways of gaining insights into a language model's internal states, so as to make its outputs more interpretable and transparent.

In this way, some have pointed to the emergence of sophisticated linguistic knowledge. For instance, Manning et al. (2020) found that BERT's word embeddings seem to encode parse tree distances by employing a structural *probe* for extracting syntactic trees from the network's internal representations. This method tries to find a distance metric between contextual word representations, such that distance between the vector representations of any two words reflects their distance in the parse tree of the sentence.

On a different note, Ribeiro et al. (2016) proposed to use model-agnostic approaches (LIME – Local Interpretable Model-Agnostic Explanations) to explain machine learning's predictions by approximating the complex model locally with a simpler, more interpretable one. This allows researchers to identify which input features most strongly influenced a given output.

Our work makes instead use of a *behavioral test*. The latter is a way of evaluating language models from the "outside", by looking at their inputs and outputs without trying to probe their hidden representations. More specifically, one uses carefully designed stimuli and observes the model's responses to them. This allows us to infer patterns of behavior that can reveal something about the model (i.e., its capabilities, limitations or biases).

Concerning our topic of interest – LLMs' understanding and sensitivity to garden-path sentences – the literature in the NLP field is rather limited. Most notably, it is worth pointing to work by Futrell et al. (2019), which investigated whether LLMs represent

syntactic states in a way comparable to human sentence processing. They considered a variety of linguistic phenomena, particularly garden-path sentences, and found for them that the models do show surprisal patterns consistent with garden-path effects. However, the reliability of the effect varied across architectures, and models were generally less robust than humans in handling these difficult constructions.

1.2. Overview of the Garden Path phenomenon:

The Garden Path theory of sentence comprehension is a *serial* parsing model, in which the sentence processor (or "parser") resolves a structural ambiguity by choosing a single parsing strategy at a time, on the basis of syntactic factors (Frazier & Rayner, 1982; Altmann & Steedman, 1988).[1] Moreover, it is a two-stage model (Frazier & Fodor, 1978), where the initial stage is constituted by the tree-structure building operation, impermeable to other types of information (such as frequency, world knowledge, plausibility), which are considered at a later stage.

Concretely, the Garden Path phenomenon is a *reanalysis* effect due to the failure of a (preferred) parsing strategy. As an example, consider sentence (1) below, a classic example of this phenomenon taken from Frazier (1987):

(1) The horse

The horse raced

The horse raced past the barn

[1] The term *parsing* refers to the process of assigning a structure to a certain sentence. Serial models of parsing, like the Garden Path model, are opposed to *parallel* models of language processing, such as the Competition model (Spivey & Tanenhaus, 1998) and the Unrestricted race model (van Gompel et al., 2000). In parallel models, multiple parses are generated in parallel, and disambiguation relies not only on grammatical factors, but also on lexical, semantic and pragmatic factors. This is why they are sometimes also referred to as *interactive*, where the interaction occurs between two streams of information – a bottom-up flow, starting from the grammar, and a top-down one, which includes higher-order information (e.g., McClelland, 1987). Thus, the two types of models – serial and parallel – essentially differ by the timing of extra-syntactic information (after syntactic computation or concurrently with it).

The horse raced past the barn **fell**

The horse [that was raced past the barn] fell

As we can see, the parser goes through the sentence and builds its structure sequentially, one constituent at a time, up to the verb *fell*. The latter is highlighted in bold because it is the disambiguating region where the garden-path effect occurs. Before its integration, the sentence "The horse raced past the barn" is interpreted by the parser as a full sentence where the verb *raced* is in the past simple tense. However, the later integration of *fell* constrains the parser to a reanalysis whereby "raced past the barn" is a reduced relative clause – and thus, *raced* is rather a past participle – and "The horse fell" is the main clause.

The preference for a certain candidate structure over another one is guided by the following tree-building internal economy principles (Frazier, 1979):

1. Minimal attachment: do not postulate any potentially unnecessary nodes.
2. Late closure: if grammatically permissible, attach new items into the clause or the phrase currently being processed.

In other words, the first principle recommends postulating as few nodes as possible, leading to the preference for a less complex structure over a more complex one (where complexity can be defined in terms of number of nodes in the tree structure). The second principle, in turn, recommends integrating as many constituents as possible into the structure that has already been built up to that point.[2]

[2] A third principle, known as the Minimal Chain Principle, was later added by De Vincenzi (1991): avoid postulating unnecessary chain members but do not delay required chain members. This principle was originally formulated in relation to structurally ambiguous sentences in Italian. An example is represented by the sentence "Chi ha chiamato?", where the pronoun *chi* (equivalent to the English *who*) can be interpreted both as a subject and as an object. According to De Vincenzi, the parser tends to resolve the ambiguity by favoring the interpretation which modifies the original argument structure the least (in this case, the subject interpretation), and which thus minimizes the long-distance dependency between the original merge position of the constituent and its overt realization (for the terminology used, see e.g., Chomsky, 1995).

We can clearly see these principles in action when comparing the two alternative parses of sentence (1) above, visualized below in **Fig.1**:[3]

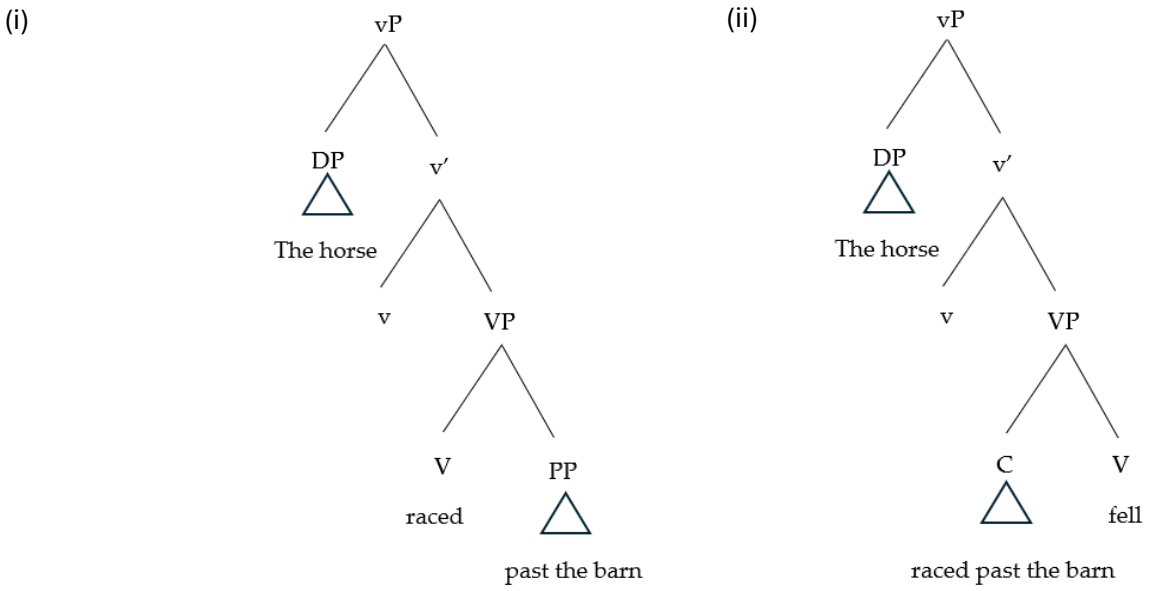

Fig.1. Sequential parsing of the sentence "The horse raced past the barn fell".

Structure (i) is initially the preferred parse because it adheres to both principles: on the one hand, the structure is as simple as possible, in compliance with Minimal attachment (whereas the structure in (ii), with the reduced relative clause, has more nodes, summarized under the complementizer label "C"); on the other hand, all the elements are easily integrated within the same already-built structure, while in option (ii) the parser has to modify and enlarge the original structure in (i).

There are different types of garden-path sentences according to a categorization used, among others, in Futrell et al. (2019). In this work, we will cover four garden-path families:

1. Main verb/relative clause (henceforth MV/RC) ambiguity, already presented in sentence (1) above and further illustrated in sentence (2), taken from Milne (1982):

   (2) The boat floated down the river sank.

[3] The syntactic trees displayed in **Fig.1** and **Fig.2** were constructed following the assumptions of work in generative linguistics (e.g., Chomsky, 1995).

The verb *floated* is initially parsed as the main verb, although it actually introduces a reduced relative clause.

2. NP/Z ambiguity: here, the verb is ambiguous between its transitive use (in which case it takes an NP object) and its intransitive use (in which case it takes a zero or null object). An example is given below in sentence (3):

   (3) As the doctor studied the textbook fell to the floor.

   The parser tends to initially interpret the verb *studied* as transitive, taking the direct NP object *the textbook*, but is later forced to a reanalysis when accounting for the rest of the sentence. In reality, this NP is the subject of the main clause, and the verb is used intransitively.

3. NP/S ambiguity: in the case, the ambiguity concerns whether the (transitive) verb takes an NP or a sentential complement. An example is constituted by sentence (4):

   (4) The director announced the cast would change.

   The verb *announced* takes not only the NP *the cast*, but the whole clause *the cast would change*.

4. Lexical categorical ambiguity: the last garden-path subtype consists in the ambiguity between different parts of speech, as exemplified in (5), where the noun *fat* may be interpreted as an adjective at first:

   (5) Fat people eat accumulates.

These four types of ambiguity will be the object of the experiment discussed in detail in the second part of this work.

2. The experiment

The second part of the paper is dedicated to the experiment: § 2.1 presents the methodology, whereas § 2.2 discusses the results.

2.1 Methodology

The first step of the project consisted in the collection of a corpus of garden-path sentences to administer to the chosen LLM, amounting to 58 sentences in total. Most of them were taken from the works of scholars who studied the Garden Path phenomenon, while some of them were hand-crafted. As already anticipated in § 1.2, four kinds of garden-path sentences were used, with the following distribution: 10 sentences of the MV/RC type, 25 of the NP/Z type, 12 of the NP/S types and 11 of the categorical ambiguity type. [4] Each sentence was then labeled with the category it belongs to, so as to check whether a significant difference could be found in the language model's behavioral response across categories.

Since the main purpose of this experiment was to assess the model's sensitivity to garden-path sentences, each sentence of the dataset was paired with a corresponding baseline stimulus, namely an unambiguous sentence minimally differing from its garden-path counterpart. See example (6) below:

(6) a. Since Jay always jogs a mile seems a short distance for him.

b. Since Jay always jogs, a mile seems a short distance for him.

Sentence a. represents the garden-path sentence (of the NP/Z type), while sentence b. constitutes the baseline. Indeed, a. initially favors the transitive reading of the verb *jogs*, until the verb *seems*, which forces a reanalysis; conversely, the use of the comma in b. immediately disambiguates the interpretation towards the intransitive use of the verb.

The final stimuli had the following format (one example per category is given):

("MV/RC", "The postman delivered junk mail threw it into the trash.", "The postman who was delivered junk mail threw it into the trash."),

("NP/Z", "The man who whistles tunes pianos.", "The man who whistles is tuning pianos"),

("NP/S", "The explorers found the South Pole was impossible to reach.", "The explorers found that the South Pole was impossible to reach."),

[4] See the Appendix for a complete list of the stimuli used.

("Categorical", "The complex houses married soldiers and their families.", "Married soldiers and their families are housed in the complex.").

Concretely, the model's performance was evaluated by calculating *perplexity* scores. The perplexity of an LLM on a test set can be defined as "the inverse probability of the test set, normalized by the number of words" (Jurafsky & Martin, 2009, p. 36), and it is formally defined as:

$$\begin{aligned} \text{perplexity}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}} \end{aligned}$$

where $W = w_1 w_2 \ldots w_N$ is the test set (Jurafsky & Martin 2009, p. 36). In other words, it is a measure of how "surprised" the model is. Given the theory of the Garden Path phenomenon presented in § 1.2, we predict that the language model's perplexity scores will be significantly higher in the garden-path condition than in the baseline condition, as an effect of the *reanalysis* elicited by the garden-path sentences. In order to verify this hypothesis, we thus calculated the model's perplexity assigned to each garden-path sentence and to its corresponding baseline sentence.

Furthermore, in order to assess the effect of the various garden-path types, we calculated delta scores - consisting in the difference in perplexities between baseline and garden-path conditions – with the idea that the higher the difference, the more surprising the model will find the garden-path sentence to be (or the baseline sentence in case of a negative difference, which we consider however unlikely). Then, we averaged the results by category and eventually compared the average surprisal delta across categories.

Even if, according to the original two-stage account discussed in § 1.2, the parsing and subsequent reanalysis of a garden-path sentence is primarily linked to syntactic principles, we can hypothesize that the degree of the reanalysis effect (and, hence, the degree of the language model's surprisal) might also be influenced by extra-syntactic factors, such as semantic plausibility, lexical information or frequency of occurrence (with the frequency of a given structure probably being inversely correlated with its computational cost). Having said that, we expect that, on average, the highest delta score will be assigned to

sentences of the MV/RC category - the "classic" garden-path sentences - due to a certain "main clause bias" (e.g., Frazier, 1979), as main clauses are structurally simpler and more common than (reduced) relative clauses. We expect them to be more surprising than both sentences of the NP/Z type and (especially) of the NP/S type, with the latter intuitively appearing to be more widespread and straightforward to interpret - probably because verbs like *to know* often take a (reduced) sentential complement. Finally, we predict sentences of the "lexical categorical ambiguity" type to be highly confusing as well. Indeed, even if their syntactic structure is generally quite simple and common, the lexical category of their constituents is not – as illustrated, for instance, by the sentence "The old man the boat", where *man* is normally not used as a verb.

Importantly, perplexity is sensitive to text length, as is evident from the definition reported above. To control for this factor, baseline and garden-path sentences were carefully constructed as minimal pairs, differing only by a few tokens. As for the comparison across categories, instead of the absolute differences (deltas) in perplexity between baseline and garden-path conditions, we calculated relative deltas by normalizing by the baseline.[5] This allowed us to examine the relative increase (or decrease, in case of a negative difference) in perplexity of the garden-path sentence with respect to its baseline.

Finally, the LLM used for this experiment was GPT-2, a language model developed by OpenAI. Like all other GPT language models, it is an *autoregressive* model, which predicts the probability of the next token by conditioning on the preceding context. It is thus suited for the task at hand, involving the use of perplexity scores, which can directly be computed from its loss. Furthermore, the unidirectional nature of its processing makes it aligned with the hypothesis of a left-to-right computation for humans (Shan & Barker, 2006, *inter alia*), as well as to the serial processing assumption of the Garden Path model, where the parser commits to a single analysis without waiting for the full sentence, assigning structure incrementally as each constituent is encountered.[6]

[5] The formula used in the code was the following: relative_delta = (gp_ppl - bl_ppl) / bl_ppl, where gp_ppl and bl_ppl stand for garden-path perplexity and baseline perplexity, respectively.

[6] In this respect, GPT models differ from bidirectional models LLMs like BERT, where the prediction of a token is conditioned upon both the left and the right context (Devlin et al., 2019).

2.2 Experiment results and analysis:

Concerning the main comparison of interest (garden-path vs. baseline sentences), the data show a clear pattern that neatly confirms our prediction outlined in § 2.1. We predicted that the model would find garden-path sentences more surprising than the corresponding baseline sentences, because of the reanalysis effect characterizing the former. Our results indeed show that, in all 58 sentence pairs, GPT-2 assigned a higher perplexity score to the garden-path sentence.

The following are some specific examples, one for each garden-path type. The tables report the perplexities assigned to both conditions, as well as the surprisal delta, while the figures illustrate the per-token log probabilities.

Example 1:

MV/RC category

Sentence type	Sentence	Perplexity	Surprisal Δ
Garden-path	*The girl told the story cried.*	559.23	+465.52
Baseline	*The girl who was told the story cried.*	93.71	

Table 1. Pair 6 (MC/RV category) - Perplexity scores and surprisal Δ.

Already looking at the table, we can observe how the model found the garden-path sentence significantly more confusing than its baseline counterpart, where the reduced relative is made explicit (perplexity scores: 559.23 vs. 93.71), as also evidenced by the surprisal Δ, given by their difference (559.23 – 93.71 = 465.52).

Looking at **Fig.2** below, we can see that the token *cried*, which represents the disambiguating region of this garden-path sentence, received the lowest log probability (around -15), which means that the model was highly surprised by it.

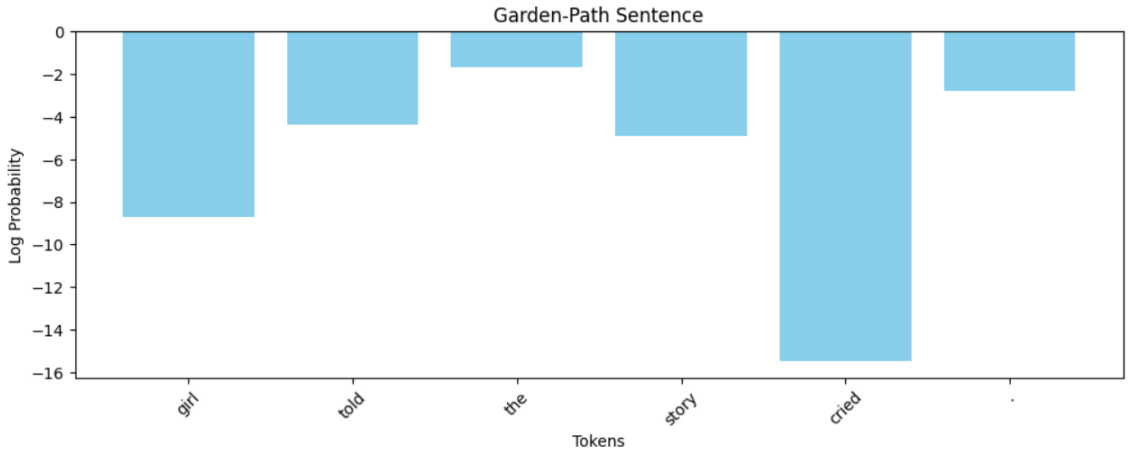

Fig. 2. Per-token log probabilities of the garden-path sentence "The girl told the story cried.".

As for the baseline sentence (**Fig.3**), the main verb *cried* is still the most surprising token (together with the noun *girl*), but the log probability assigned to it is much higher (around -9).

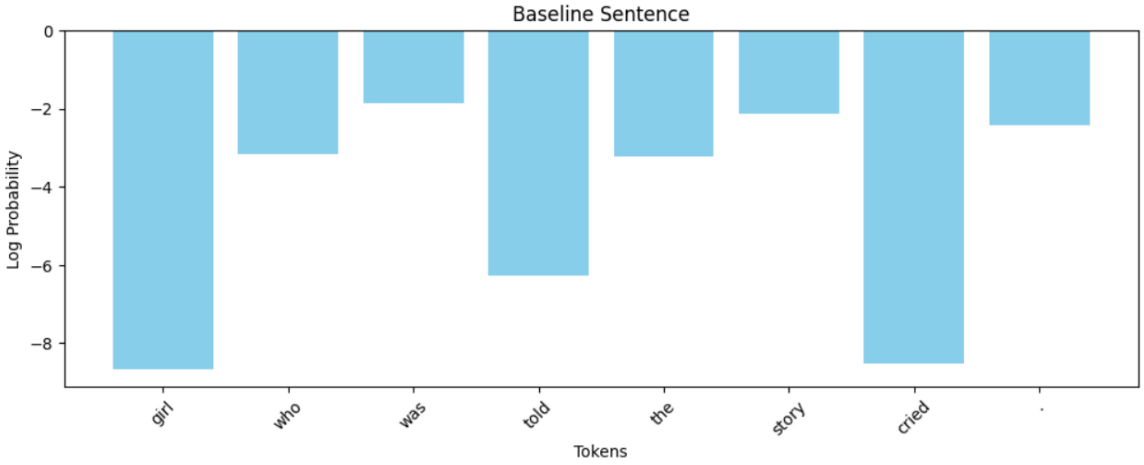

Fig.3. Per-token log probabilities of the baseline sentence "The girl who was told the story cried.".

Example 2:

NP/S category

Sentence type	Sentence	Perplexity	Surprisal Δ
Garden-path	*She told her daughter a dream could be achieved.*	78.57	+29.65
Baseline	*She told her daughter that a dream could be achieved.*	48.93	

Table 2. Pair 47 (NP/S category) - Perplexity scores and surprisal Δ.

In this example from the NP/S category, the overall perplexity scores (78.57 and 48.93) and their delta (+29.65) are much lower than the previous case. Nonetheless, the same trend can be observed: the model finds the garden-path sentence more confusing than the baseline. More specifically, **Fig.4** shows that the most surprising token is *could* (with a log probability of around -7), which is, again, the disambiguating region.

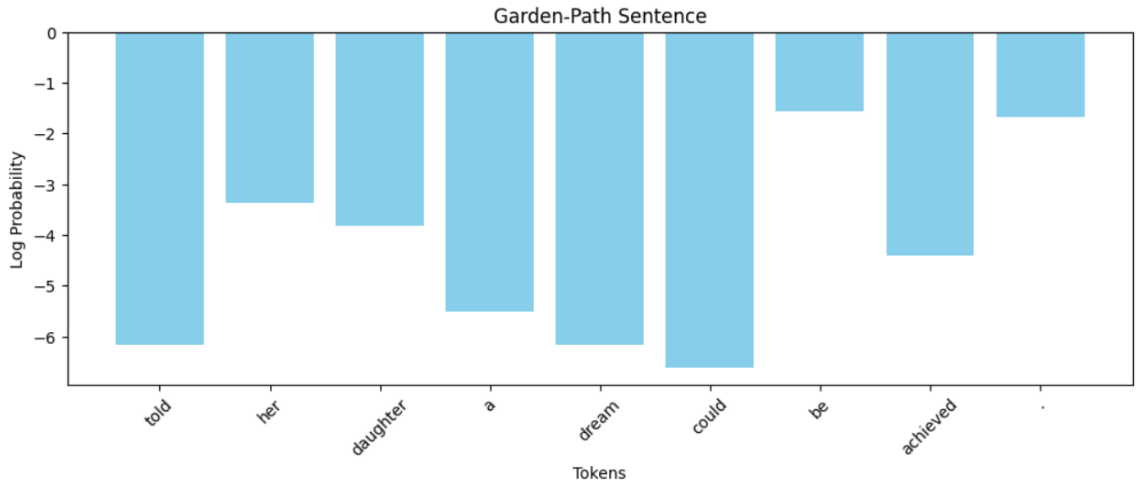

Fig.4. Per-token log probabilities of the garden-path sentence "She told her daughter a dream could be achieved.".

The same token scores a log probability of -4.5 in the baseline, where the sentential complement of the verb *told* is made explicit by the use of the complementizer *that*. This is shown in **Fig.5** below:

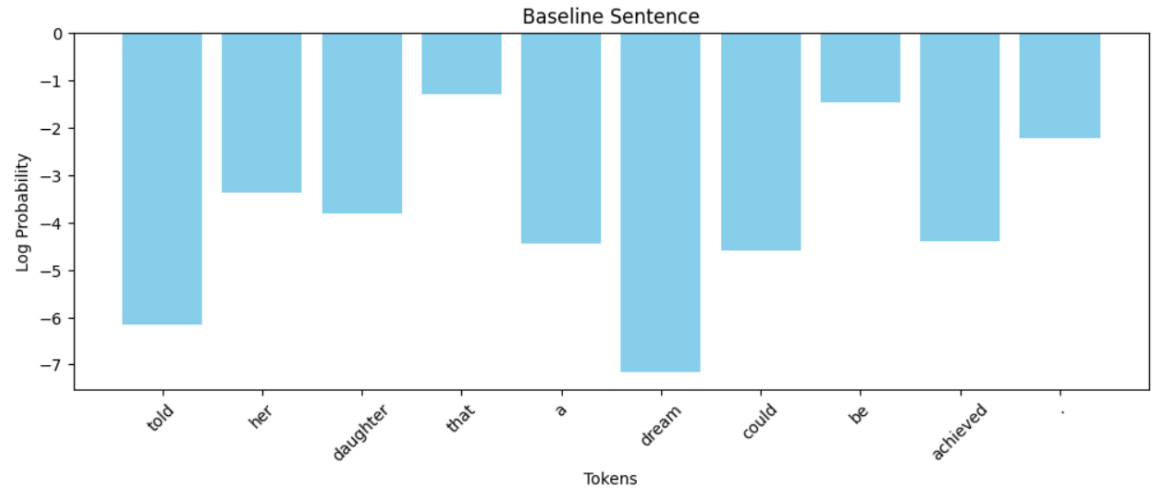

Fig. 5. Per-token log probabilities of the baseline sentence "She told her daughter that a dream could be achieved.".

Example 3:

Lexical categorical ambiguity

Sentence type	Sentence	Perplexity	Surprisal Δ
Garden-path	*The building blocks the sun faded are red.*	718.84	+328.76
Baseline	*The building blocks that the sun faded are red.*	390.08	

Table 3. Pair 49 (Categorical) - Perplexity scores and surprisal Δ.

Here, we can notice that the perplexity score is quite high also in the baseline sentence, even if once again the overall pattern is confirmed, with a positive difference of 328.76 between garden-path and baseline conditions. **Fig.6** and **Fig.7** provide a more detailed account:

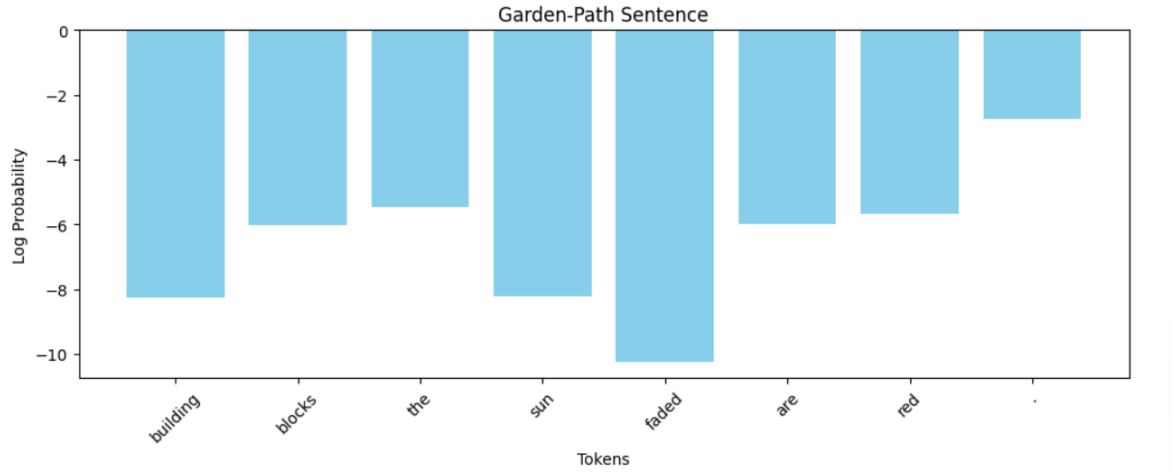

Fig.6. Per-token log probabilities of the garden-path sentence "The building blocks the sun faded are red.".

In **Fig.6**, the most confusing token is *faded* (with a log probability of around -10), which corresponds to the site of the reanalysis effect typical of garden-path sentences. This token surprisingly received an even lower log probability in the baseline condition (**Fig.7** below),

around -11, even though this result is most likely explained by the unusual use of *faded* as a transitive verb.

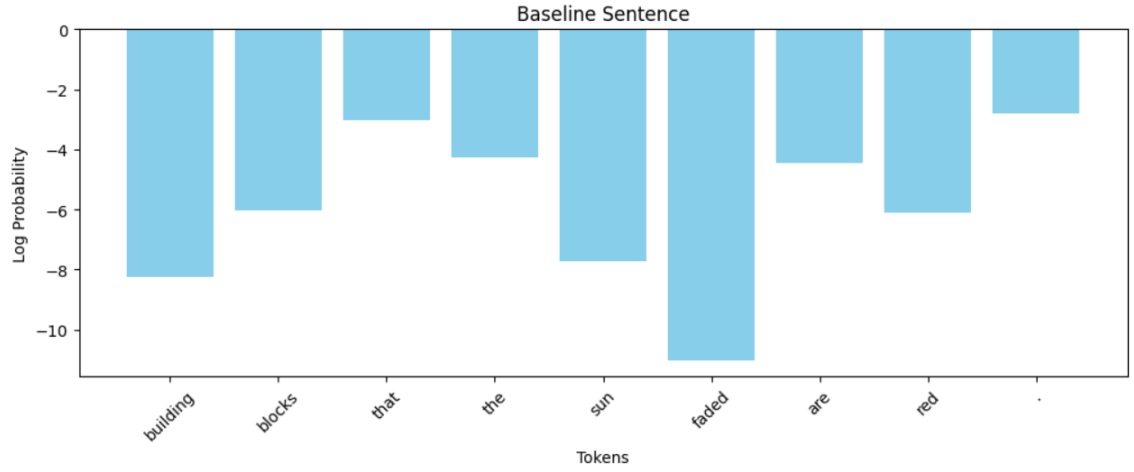

Fig.7. Per-token log probabilities of the baseline sentence "The building blocks that the sun faded are red".

Example 4:

NP/Z category

Sentence type	Sentence	Perplexity	Surprisal Δ
Garden-path	*While the lawyer studied the contract lay on the desk.*	174.42	+32.05
Baseline	*While the lawyer studied, the contract lay on the desk.*	142.37	

Table 4. Pair 25 (NP/Z) - Perplexity scores and surprisal Δ.

The last example is from the NP/Z ambiguity type. From **Table 4**, we can observe a relatively small difference in perplexity between the two conditions, namely +32.05, which is positive due to the garden-path perplexity score being higher.

Looking at **Fig.8**, we can see that the token *lay*, which triggers the reanalysis effect in the sentence at hand, is by far the most surprising to the model, with a log probability of around -12. By contrast, in the baseline condition, the log probability of this token increases to around -9, in line with other tokens of the sentence like *lawyer* and *studied*.

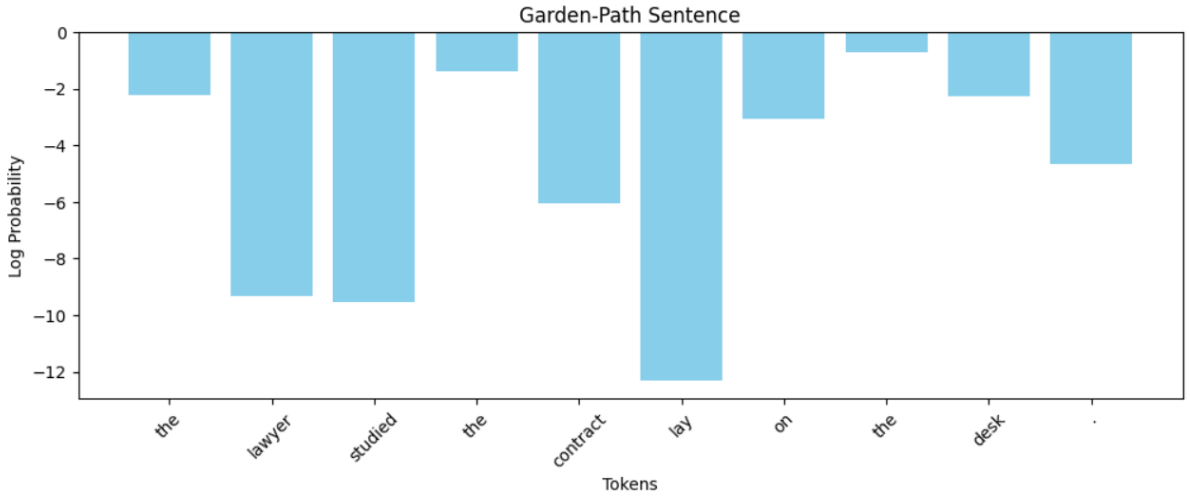

Fig.8. Per-token log probabilities of the garden-path sentence "While the lawyer studied the contract lay on the desk.".

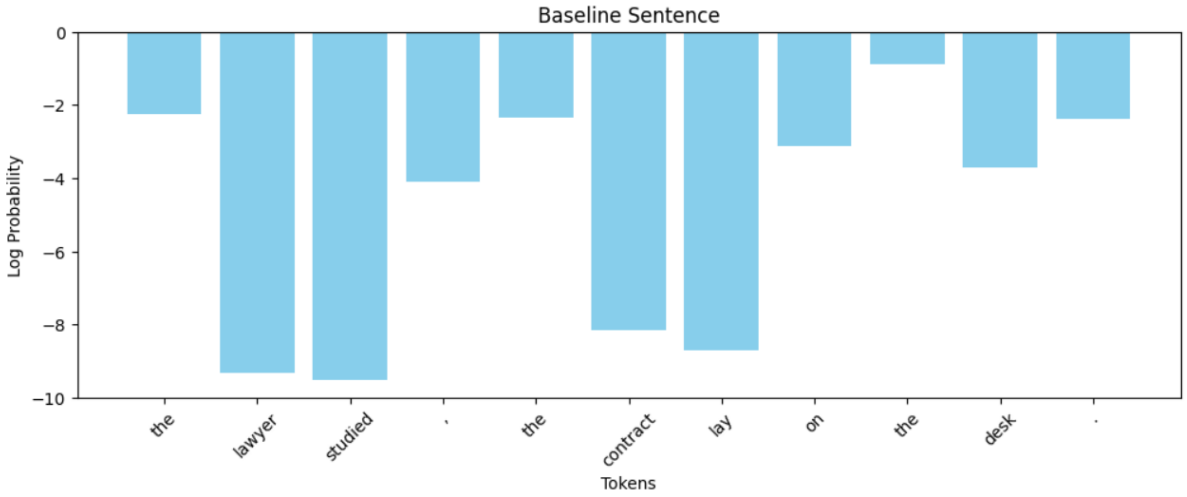

Fig.9. Per-token log probabilities of the baseline sentence "While the lawyer studied, the contract lay on the desk.".

Finally, we calculated the relative increase in perplexity of garden-path sentences with respect to baselines for a comparison across categories.[7] The results are can be visualized in **Fig.10**:

[7] Since garden-path sentences were systematically higher in perplexity than their baselines, no relative *decrease* was found.

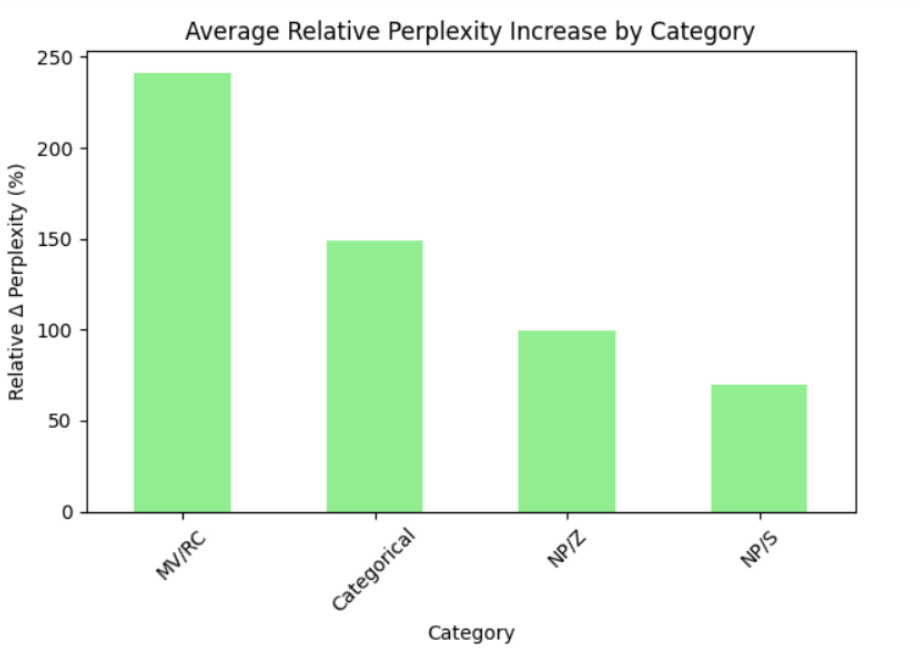

Fig.10. Average relative delta (as %) by garden-path category.

Overall, these results are aligned with the predictions we made in § 2.1. First, as we expected, and as could already be guessed from Example 2 above, sentences of the NP/S type present the smallest increase in perplexity with respect to their baselines (around 70%). This is probably due to the structure of these sentences being quite common. While frequency of occurrence does not cancel out the reanalysis effect, which is indeed present, the magnitude of such effect is not comparable to that of other categories.

The percentage is slightly higher for the NP/Z category (around 100%) and clearly higher for the Lexical categorical ambiguity type (around 140%). As already mentioned in § 2.1, the lexical categories of the constituents can be quite unusual in this garden-path subtype, and thus the reanalysis effect here is essentially triggered by the failure of a parsing strategy more aligned with lexical expectations.

Finally, as expected, the MV/RC type presents the highest increase in perplexity of the garden-path condition compared to the baseline. As suggested in § 2.1, this is likely explainable by the preference for the main clause over reduced relative clauses.

Conclusion

The purpose of this paper was to contribute to work in interpretability in the field of NLP. With this aim, we proposed a behavioral test to assess GPT-2's sensitivity to garden-path sentences. First, we collected a dataset of garden-path-baseline sentence pairs, sorted into four different garden-path categories. Then, we calculated the model's perplexity scores assigned to each garden-path sentence and its baseline counterpart. We had the expectation that such scores would align with the human processing difficulties associated with garden-path sentences, as an effect of the failure of a preferred analysis and subsequent choice of an alternative parsing strategy (the so-called *reanalysis* effect). We found indeed a systematic tendency of the model to assign higher perplexities to the garden-path sentences, reflecting higher surprisal.

At the same time, we investigated the possible differences in perplexity across the different categories of garden-path sentences. The results indicate that sentences of the MV/RC type are the most perplexing, followed by Categorical ambiguity sentences, NP/Z sentences and, lastly, NP/S sentences (by far the least surprising ones for the model).

This work is not intended as the proposal of a definitive method, but rather as an initial steppingstone which can benefit from future improvements. The following are some suggestions for further research in this direction: since the Garden Path is a model of human sentence processing, we think that, for completeness, the machine experiment should be compared with data derived from a similar human experiment. The latter could be, for instance, an EEG experiment measuring language-related ERPs.[8] In particular, the N400 component has been traditionally associated with sematic oddness (e.g., Kutas & Hillyard, 1980; Lindborg et al., 2023, *inter alia*), and as such it could provide relevant information on the processing of garden-path sentences. We expect this component to be elicited on the disambiguating region of a garden-path sentence.

Moreover, it would be interesting to see if humans respond differently to different types of garden-path sentence, and whether they would reproduce the surprisal ranking observed here for GPT-2 for the examined categories.

[8] Event-related potentials (see Sutton et al., 1965, *inter alia*).

REFERENCES

[1] Altmann, G., Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30*(3), 191–238. https://doi.org/10.1016/0010-0277(88)90020-0

[2] Chomsky, N. (1995). *The Minimalist Program.* Cambridge, M.A.: MIT Press.

[3] de Vincenzi, M. (1991). *Syntactic parsing strategies in Italian: The minimal chain principle*. Studies in Theoretical Psycholinguistics, Vol. 12. Springer. https://doi.org/10.1007/978-_94-011-3184-1

[4] Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In Association for Computational Linguistics (ACL). 4171–4186.

[5] Du, J.-L., & Yu, P.-F. (2010). Towards natural language processing: A well-formed substring table approach to understanding garden path sentences. In *2010 2nd International Workshop on Database Technology and Applications* (pp. 1–5). IEEE. https://doi.org/10.1109/DBTA.2010.5659102

[6] Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of 'garden-path' sentences. *Journal of Memory and Language, 30*, 725–745.

[7] Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research, 30*(1), 3–20. https://doi.org/10.1023/A:1005290706460

[8] Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition, 6*(4), 291–325. https://doi.org/10.1016/0010-0277(78)90002-1

[9] Frazier, L (1979), on comprehending sentences: syntactic parsing strategies. *Doctoral Dissertations*. AAI7914150. https://digitalcommons.lib.uconn.edu/dissertations/AAI7914150

[10] Frazier, L., and Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences.

Cognitive Psychology, 14(2), 178–210. https://doi.org/10.1016/0010-0285(82)90008-1

[11] Frazier, L. (1987). Theories of sentence processing. In J. L. Garfield (Ed.), *Modularity in knowledge representation and natural-language understanding* (pp. 291–307). The MIT Press.

[12] Futrell, R., Wilcox, E. Morita, T., Qian, P., Ballesteros, M., and Levy, R. (2019). Neural Language Models as Psycholinguistic Subjects: Representations of Syntactic State. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 32–42. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1004.

[13] Jurafsky, D., and Martin, J.H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Prentice Hall.

[14] Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203–205. https://doi.org/10.1126/science.7350657

[15] Lindborg, A., Musiolek, L., Ostwald, D., & Rabovsky, M. (2023). Semantic surprise predicts the N400 brain potential. *Neuropsychologia: Reports, 3*, 100161. https://doi.org/10.1016/j.ynirp.2023.100161

[15] Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., and Levy, O. (2020). *Emergent linguistic structure in artificial neural networks trained by self-supervision*. Proceedings of the National Academy of Sciences 117(48). 30046–30054.

[16] McClelland, J. L. (1987). The case for interactionism in language processing. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 1–36). Hillsdale, NJ: Lawrence Erlbaum Associates.

[17] Milne, R. W. (1982). Predicting garden path sentences. *Cognitive Science, 6*(4), 349–373.

https://doi.org/10.1207/s15516709cog0604_3

[18] Patson, N. D., Darowski, E. S., Moon, N., & Ferreira, F. (2009). Lingering misinterpretations in garden-path sentences: Evidence from a paraphrasing task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(1), 280–285. https://doi.org/10.1037/a0014276

[18] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. In ICML Workshop on Human Interpretability in Machine Learning (WHI).

[19] Sarti, G. (2020). *Interpreting neural language models for linguistic complexity assessment* (Master's thesis, University of Trieste). SISSA - Scuola Internazionale Superiore di Studi Avanzati.
https://github.com/gsarti/interpreting-complexity

[20] Shan, C.-C., & Barker, C. (2006). Explaining crossover and superiority as left-to-right evaluation. *Linguistics and Philosophy*, *29*(1), 91–134. https://doi.org/10.1007/s10988-005-6580-7

[21] Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic Ambiguity Resolution in Discourse: Modeling the Effects of Referential Context and Lexical Frequency. Journal of Experimental Psychology: Learning Memory Cognition, 24, 1521-43. https://doi.org/10.1037/0278-7393.24.6.1521

[22] Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science, 150*(3700), 1187–1188. https://doi.org/10.1126/science.150.3700.1187

[23] van Gompel, R. P. G., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*. https://doi.org/10.1016/j.jml.2004.11.003

[24] Yaman, S. (2023). *Processing and interpretation of garden-path sentences in L2 speakers of*

English (Master's thesis, Middle East Technical University, Graduate School of Social Sciences).

APPENDIX

Corpus of garden-path – baseline sentence pairs, sorted by category (source indicated):

MV/RC ambiguity:

- The horse raced past the barn fell.
  The horse that raced past the barn fell.
- Vaccine trials halted after a patient fell ill restart. (from Sarti, 2020)
  Vaccine trials, which were halted after a patient fell ill, restart.
- The florist sent the flowers was rather pleased.
  The florist who was sent the flowers was rather pleased.
- The performer sent the flowers was rather pleased.
  The performer who was sent the flowers was rather pleased.
- The girl told the story cried.
  The girl who was told the story cried.
- The cotton clothing is made of grows in Mississippi.
  The cotton that clothing is made of grows in Mississippi. (from Du & Yu, 2010)
- The boat floated down the river sank. (from Milne, 1982)
  The boat that floated down the river sank.
- The postman delivered junk mail threw it into the trash. (from Milne, 1982)
  The postman who was delivered junk mail threw it into the trash.
- The man gifted a ring played the piano.
  The man who was gifted a ring played the piano.
- The old man told the son the story complained to his son.
  The old man who was told the story complained to his son.

Np/Z ambiguity:

- As the criminal shot the woman yelled at the top of her lungs. (from Sarti, 2020)
  As the criminal shot, the woman yelled at the top of her lungs.
- As the doctor studied the textbook fell to the floor.
  As the doctor studied, the textbook fell to the floor.
- Since Jay always jogs a mile seems a short distance to him. (Frazier & Rayner, 1982)
  Since Jay always jogs, a mile seems a short distance to him.
- When the dog scratched the vet with his new assistant took off the muzzle.
  When the dog struggled the vet with his new assistant took off the muzzle. (Futrell et al., 2019).
- The man who whistles tunes pianos.
  The man who whistles is tuning pianos.

- Because Bill drinks wine is never kept in the house. (from Ferreira & Henderson, 1991)
  Because Bill drinks, wine is never kept in the house.
- When the boys strike the dog kills. (from Ferreira & Henderson, 1991)
  When the boys strike, the dog kills.
- After the Martians invaded the town was evacuated. (from Ferreira & Henderson, 1991)
  After the Martians invaded, the town was evacuated.
- After the Martians invaded the town the city bordered the people were evacuated.
  After the Martians invaded the town that the city bordered, the people were evacuated.
- While Bill hunted the deer ran into the woods. (Source: Ferreira et al., 2001)
  While Bill hunted, the deer ran into the woods.
- While Anna bathed the kid played in the crib.
  While Anna bathed, the kid played in the crib.
- The man who hunts ducks out on weekends.
  The hunter ducks out on weekends.
- As the woman photographed a bird sang beautifully.
  As the woman photographed, a bird sang beautifully.
- As the artist was painting the portrait fell to the ground.
  As the artist was painting, the portrait fell to the ground.
- While my mum was cooking pasta boiled in the pot.
  While my mum was cooking, pasta boiled in the pot.
- While the nurse shaved the patient who was tired and weak watched TV. (from Patson et al., 2009)
  While the nurse shaved, the patient who was tired and weak watched TV.
- While the orchestra performed the symphony played on the radio. (from Patson et al., 2009)
  While the orchestra performed, the symphony played on the radio.
- While the lawyer studied the contract lay on the desk.
  While the lawyer studied, the contract lay on the desk.
- While Rick drove the car that was red and dusty veered into a ditch. (from Patson et al., 2009)
  While Rick drove, the car that was red and dusty veered into a ditch.
- While the secretary typed the memo that was clear and concise neared completion. (from Patson et al., 2009)
  While the secretary typed, the memo that was clear and concise neared completion.
- As John whittled the wooden stick broke in half.
  As John whittled, the wooden stick broke in half.
- While the mother calmed down the children who were irritable and tired sat on the bed.

While the mother calmed down, the children who were irritable and tired sat on the bed. (from Patson et al., 2009)

- While the farmer steered the tractor pulled the plough.
  While the farmer steered, the tractor pulled the plough.
- While the chimps groomed the baboons that were large and hairy sat in the grass.
  While the chimps groomed, the baboons that were large and hairy sat in the grass. (from Patson et al., 2009)
- Without greeting the man ran away.
  Without greeting, the man ran away.

NP/S ambiguity:

- The explorers found the South Pole was impossible to reach.
  The explorers found that the South Pole was impossible to reach. (from Yaman, 2023)
- The teacher knew the student had been struggling for a while.
  The teacher knew that the student had been struggling for a while.
- The fan finally saw his favorite baseball player had taken the field.
  The fan finally saw that his favorite baseball player had taken the field.
- The director announced the cast would change.
  The director announced that the cast would change.
- The mum saw her child had finished eating.
  The mum saw that her child had finished eating.
- The girl noticed a spot was staining the wall.
  The girl noticed that a spot was staining the wall.
- Henry eventually realized his mistake could not be fixed.
  Henry eventually realized that his mistake could not be fixed.
- The professor warned the student was late.
  The professor warned that the student was late.
- The journalist read the magazine had been closed.
  The journalist read that the magazine had been closed.
- Gyada knows the old man is being honest.
  Gyada knows that the old man is being honest.
- She told her daughter a dream could be achieved. (from Yaman, 2023)
  She told her daughter that a dream could be achieved.
- The employees read the contract would expire very soon.
  The employees read that the contract would expire very soon.

Lexical categorical ambiguity:

- Fat people eat accumulates.
  Fat that people eat accumulates.
- The old man the boat. (from Du & Yu, 2010).

It is the old who man the boat.

- The complex houses married soldiers and their families.
  Married soldiers and their families are housed in the complex.
- The granite rocks by the seashore with the waves. (from Milne, 1982)
  It is the granite that rocks by the seashore with the waves.
- The prime number few. (from Milne, 1982)
  Prime individuals count few members.
- The sour drink from the ocean.
  Sour people drink from the ocean.
- The tired run late at night.
  The tired people run late at night.
- The experienced coach the football team.
  The experienced players coach the football team.
- The young soldier the barracks at dawn.
  The young people soldier the barracks at dawn.
- The joyful band the parade.
  The joyful guys band the parade.
- The building blocks the sun faded are red. (from Milne, 1982)
  The building blocks that the sun faded are red.