Extracting structure from an LLM - how to improve on surprisal-based models of human language processing

Daphne Wang
Quandela

Mehrnoosh Sadrzadeh
University College London

Miloš Stanojević
University College London

Wing-Yee Chow and Richard Breheny
University College London

Abstract

Prediction and reanalysis are considered two key processes that underlie humans' capacity to comprehend language in real time. Computational models capture it using Large Language Models (LLMs) and a statistical measure known as 'surprisal'. Despite successes of LLMs, surprisal-based models face challenges when it comes to sentences requiring reanalysis due to pervasive temporary structural ambiguities, such as garden path sentences. We ask whether structural information can be extracted from LLM's and develop a model that integrates it with their learnt statistics. When applied to a dataset of garden path sentences, the model achieved a significantly higher correlation with human reading times than surprisal. It also provided a better prediction of the garden path effect and could distinguish between sentence types with different levels of difficulty.

1 Introduction

Psycholinguistic research has established that prediction over sequences of words and their most common grammatical structures play an important role in human language comprehension (Frazier, 1987a; Pritchett, 1988). In this process, our ability to anticipate upcoming language enables our brain to process complex linguistic information more efficiently. Predictive processes are understood to sometimes lead to difficulties in processing. There is now a large body of empirical evidence that the difficulty associated with processing a phrase is modulated by both its structural properties and the predictability of its sequence of words (Sturt et al., 1999; Pickering and Frisson, 2001).

Current computational models capture prediction using Large Language Models (LLMs) and a statistical measure known as 'surprisal' (Hale, 2006; Levy, 2008). LLM's do not learn an explicit account of structure and it has been shown that they underestimate difficulties humans overcome due to temporary structural ambiguities in phenomena such as garden path sentences (van Schijndel and Linzen, 2021; Arehalli et al., 2022; Huang et al., 2023). We conjecture that structural information of the appropriate kind can be extracted from LLM's and develop a novel model of human language processing to combine this structure with the inherent statistics learnt by an LLM.

Our model is based on presheaves: a mathematical framework that combines the structure present in a phenomena with its probabilistic data. The framework leads to a structure-aware measure that computes the distance between actual and predicted probability distributions, while taking both statistical data (over the completion of subphrases) and their grammatical structures into account. We use this measure to model human reading times. The performance of the model is evaluated on a dataset of garden path sentences (Sturt et al., 1999) and compared with the surprisal-based models, where it did significantly better and was able to distinguish between easy and hard garden path effects; here, surprisal has so dar failed to produce reliable results (van Schijndel and Linzen, 2021).

2 Related Work

Research on temporal ambiguities goes back to the work of Bever in 1970's (Bever, 1970). These occur in so called "garden path" sentences (a term he coined) where local subphrases that have certain structural ambiguities whose favoured resolution leads to a local structure which conflicts with the correct overall meaning of the sentence. Examples of such sentences are:

(1a) (NP/S) The traveller heard the clock had woken everybody up.

(2a) (NP/Z) Before the traveller packed the clock had woken everybody up.

Bever argued that humans stall when understanding these sentences and used this as a basis for

Proceedings of the 31st International Conference on Computational Linguistics, pages 4938–4944
January 19–24, 2025. ©2025 Association for Computational Linguistics

	Regions			
NP/S	The foreign traveller	heard the loud clock	**had woken everybody up**	in the youth hostel.
NP/S (control)	The foreign traveller	heard that the loud clock	**had woken everybody up**	in the youth hostel.
NP/Z	Before the traveller	packed the loud clock	**had woken everybody up**	in the youth hostel.
NP/Z (control)	Before the traveller	packed, the loud clock	**had woken everybody up**	in the youth hostel.

Table 1: Example of different regions of Sturt et. al. dataset. The critical regions are in bold.

developing a theory of human language processing based on linguistic structure. Psycholinguistic experiments on human reading times followed suit, showing that garden path sentences take longer to comprehend than their unambiguous controls (Frazier, 1979, 1987b; Frazier and Rayner, 1990). It was later revealed that the structure of a garden path sentence affects the degree of difficulty of its comprehension (Pritchett, 1988; Ferreira and Henderson, 1990; Garnsey et al., 1997). Different types of structure were considered but the structural complexities of the parse tree (Pritchett, 1988), gathered more empirical evidence (Sturt et al., 1999). According to this theory, (2a) is harder to understand than (1a), since it has a verb that may or may not have an NP complement. These types of sentences are called NP/Z (Z for Zero complement). This is in contrast to (1a) where the verb will always have a complement, but the difficulty lies in the fact that the role of this complement is ambiguous: it can be a noun phrase (NP) or a sentence (S). These kinds of sentences are called NP/S.

More recently, an information theoretic quantity called ‘surprisal’ was used to measure lexical expectation and predict human reading times (Hale, 2001). Surprisal measures the degree of unpredictability of a word w given its prefix context $w_1 \cdots w_n$ and is computed as follows:

$$SP(w_n|w_1...w_{n-1}) = -\log(P(w_n|w_1...w_{n-1}))$$

Experimental evidence for Hale’s theory was provided by Levy (Levy, 2008) but focused on naturalistic materials, such as newspaper articles. Garden path sentences were studied by Linzen and colleagues (Schijndel and Linzen, 2018; van Schijndel and Linzen, 2021; Prasad and Linzen, 2021), but it was discovered that while surprisal can to some extent capture the increase in their processing difficulties (with some underestimation), it cannot distinguish between the different reading times of NP/Z vs NP/S sentences, whereas human reading times significantly increased.

Such underestimation of processing difficulty may indicate that either surprisal is an inadequate measure of linguistic predictability, and/or factors beyond prediction affect language processing difficulty. The model of this paper overcomes both of these potential problems.

The theory of sheaves was originally developed to model partial differential equations (Grothendieck, 1957), but soon after it was also applied to provide a general model of logical reasoning (Lane et al., 1992). Sheaves and presheaves have been used to combine the structure and data coming from physical experiments (Abramsky and Brandenburger, 2011; Barbosa, 2014), signal processing (Robinson, 2017), graph neural networks (Bodnar et al., 2022), and natural language processing (Wang et al., 2021a,b; Lo et al., 2022; Huntsman et al., 2024; Philips, 2019; Bradley et al., 2022). The use of sheaves for modelling garden path sentences is, however, novel. The only existing work is Wang et al. (2024), where the positive effects of measures similar to ours are studied, but the underlying dataset was focused on detecting the ability of the model in predicting plausibility.

3 Methodology and Experiments

Mathematical Model. Our mathematical model consisting of a base topology $\mathcal{X} = (X, \leq)$ over all incremental subphrases of a sentence, ordered by inclusion. Suppose the vocabulary of a sentence ϕ is the set σ, then the set of all phrases over it is the monoid $X = \sigma^*$ and for $m_1, m_2, \cdots \in \sigma$, we have

$$m_1 \leq m_1 m_2$$

Over this we defined a “presheaf” map P that sends each phrase $m \in \sigma$ to the set $P(m)$ of its *data*. The elements of $P(m)$ are called *sections* over m. They correspond to the set of possible grammatical structures of the subphrase m. We work with (unlabelled) dependency grammars, so every grammatical structure can be expressed as a function

$$s\colon m \to O$$

In the above, O is the set of all head positions and hence $s(w)$ is the head of w; for details see (Wang, 2024).

	$\mathbf{IF}_{min}$	$\mathbf{IF}_{JS}$	SP	Human
garden path	1110.57 ± 130	1153.01 ± 145	884.98 ± 172	1106
control	993.80 ± 116	965.23 ± 119	897.80 ± 168	862
p-value	4.59×10^{-7}	1.35×10^{-12}	0.672	

Table 2: Predicted and actual reading times in the critical region (ms)

	$\mathbf{IF}_{min}$	$\mathbf{IF}_{JS}$	SP	Human
NP/S	96.07 ± 79	163.60 ± 119.29	2.05 ± 42	87
NP/Z	137.48 ± 76	211.97 ± 99	-27.69 ± 51	400
p-value	0.0396	0.0873	0.0148	

Table 3: Predicted garden path effect for NP/S and NP/Z sentences.

For each subphrase $m \in \sigma^*$, its probability distributions are the following maps:

$$d_m \colon O(m) \to \mathbf{R}^+$$

These are measured over the set of grammatical structures $O(m)$ and as a result we require the following equation to hold:

$$\sum_{o \in O(m)} d_m(o) = 1$$

We define a structure-aware measure of difficulty called *Incompatibility Fraction* (IF) to compute the distance between the actual and the expected probability distributions of the grammatical structures of subphrases. Here we have a few options:

Divergence from average. Given two subphrases $m_1, m_1m_2 \in \sigma^*$, suppose the distributions over their grammatical structures are d_{m_1} and $d_{m_1m_2}$ and their average is as defined below:

$$M = 1/2(d_{m_1} + d_{m_1m_2}|_{m_1})$$

In the above, $d_{m_1m_2}|_{m_1}$ is the restriction of the distribution over m_1m_2 to its prefix m_1. The divergence from average is computed using the *Jensen-Shannon* metric:

$$\begin{aligned} &\mathbf{IF_{JS}}(m_1, m_1m_2) := \\ &1/2\mathbf{KL}(d_{m_1}||M) + 1/2\mathbf{KL}(d_{m_1m_2}|_{m_1}||M) \end{aligned}$$

KL is the Kullback–Leibler or KL-divergence, i.e. the formula below:

$$\mathbf{KL}(\mu||\nu) = \sum_o \mu(o) \log \frac{\mu(o)}{\nu(o)}$$

We could have used **KL** directly as a measure between probability distributions. However, since it is not defined in cases when support of μ contains elements not present in ν, we opted for the JS-divergence instead.

Distance from overlap. The overlap is computed by restricting the probability of the larger subphrase to the prefix, i.e. $d_{m_1m_2}|_{m_1}$ and taking their minimum as follows:

$$\sum_o \min(d_{m_1m_2}|_{m_1}(o), d_{m_1}(o))$$

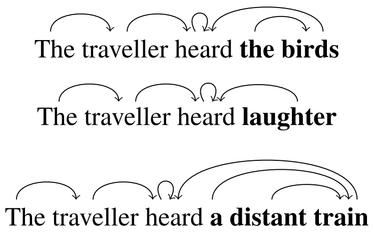
The restriction computes the marginals of the probabilities, i.e. of m_1m_2 when restricted to m_1. The distance is this overlap minus 1, defined below:

$$\begin{aligned} &\mathbf{IF_{min}}(m_1, m_1m_2) := \\ &1 - \textstyle\sum_o \min(d_{m_1m_2}|_{m_1}(o), d_{m_1}(o)) \end{aligned}$$

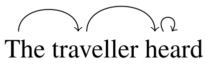
Data Collection In order to obtain the grammatical structures of the subphrases of a sentence and their probability distributions, we use two deep-learning models. These are (i) the 2nd edition of the freely accessible LLMs released by Open AI, known as `GPT-2`, which uses the state-of-the-art transformer deep neural network architecture, and (ii) The most recent release of the open-source dependency parser spaCy which also uses a state-of-the-art transformer architecture to perform a range of syntactic tasks. In the computational stage, we pass the subphrases of our garden path sentences through `GPT-2` to predict continuations. These are passed to spaCy to obtain the grammatical structures of the subphrases and their continuations. For example, below are the different continuations of (1a) and their parses:

	$RT_{\mathbf{IF}_{\text{measure}}}$(region)
min	$333.85 \sum_{w \in \text{region}} \mathsf{IF}_{min}(w) + 702.81\ ms$
JS	$455.82 \sum_{w \in \text{region}} \mathsf{IF}_{JS}(w) + 720.09\ ms$
SP	$2.23 \sum_{w \in \text{region}} SP(w) + 178.68\ ms$

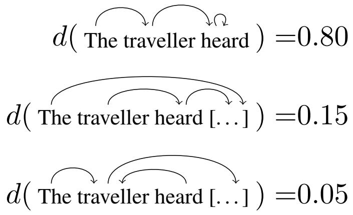
Table 4: Regression equations between **IF** and human reading times.


The traveller heard **the birds**
The traveller heard **laughter**
The traveller heard **a distant train**

The above have the same partial parse when restricted to the prefix "The traveller heard",



We sampled 1000 continuations and calculated the probability distributions of each partial tree by normalizing. Some example distributions are provided below:


d(The traveller heard) =0.80
d(The traveller heard [...]) =0.15
d(The traveller heard [...]) =0.05

After examining some of the samples, we found out that spaCy did correctly parse the garden path sentences and GPT-2 did produce garden path sentences; e.g. out of 1000 samples, the prefix "The faithful employees understood the technical contract" produced 79 sentences sharing the garden path grammatical structure, such as "The faithful employees understood the technical contract would help them do their job better". This shows that garden path sentences are not "unnatural", but trigger a non-trivial syntactic processing.

The estimation of the difference between what is observed in one stage and what was expected by the language model at the previous stage is computed by measuring the distances between the distributions obtained at the different stages via our **IF** measures, these are recorded in our dataset which is avalable online https://github.com/wangdaphne/incompatibility-fraction/tree/main. .

Experiments We work with the dataset of Sturt et al.(Sturt et al., 1999). This dataset has 24 participants and consists of 32 ambiguous NP/S, and 32 ambiguous NP/Z garden path sentences. Each sentence is paired with an unambiguous control, leading to a total of 128 sentences. The sentences are divided into 4 regions, see Table 1. The dataset contains self-paced reading times for each region.

We take participants' reading times to reflect processing difficulty and examine the extent to which $\mathbf{IF_{min}}$ and $\mathbf{IF_{JS}}$ correlate with human reading times by training a regression model. The regression coefficients are then used to (i) predict reading times and (ii) compute the garden path effect for NP/S vs NP/Z sentences. The results are compared with the same effects in humans. The tests are repeated with surprisal and the results compared.

4 Results and Analysis

The regression equations between our **IF** measures and human reading times (RT) are given in Table 4. They indicate strong positive correlation between both IF's and human reading times, see Table 5:

The Pearson's ρ coefficients are high and their p-values statistically significant. In contrast, surprisal SP has a lower coefficient and is not statistically significant. Bootstrapping (Koehn, 2004) showed that the two **IF**'s significantly outperform SP (p-values $< 10^{-140}$).

Since longer text segments are read slower than shorter ones, we expect that normalising w.r.t. token's length would improve our predictions. On the other hand, preliminary examinations did not show much improvement. We believe this is since the current dataset was carefully designed such that all sentences (as well as regions therein) were similar in size.

Predicting the reading time We used the regression equations to predict the reading times of the critical regions of the garden path sentences and their controls, see Table 2. Both **IF** measures predicted the times that are very close to those in humans (117 and 183 vs 244 ms). In contrast,

	$\mathbf{IF}_{min}$	$\mathbf{IF}_{JS}$	SP
ρ	0.8744	0.8805	0.5536
p-value	1.99×10^{-4}	1.57×10^{-4}	0.062

Table 5: Correlation and p-values between **IF** and human reading times.

surprisal predicted times that were both very close to each other (for garden path sentences and their controls) and both only close to the human reading times of the controls. The **IF** results were highly significant, but SP was not.

Distinguishing NP/S from NP/Z The IF measures were able to distinguish between the predicted garden path effects for NP/S and NP/Z sentences, see Table 3. A "garden path effect" is the difference between reading times for garden path sentences and their controls. Here, $\mathbf{IF_{JS}}$ provided a better difference than $\mathbf{IF_{min}}$ (48 ms in contrast to 41 ms). This difference was, however less than that for humans (313 ms). On the contrary, surprisal predicted the wrong trend, i.e. a higher reading time for NP/S than for NP/Z and a negative difference (-25 ms). As the p-values show, all these differences were produced with a high confidence.

5 Conclusion and Future Work

We introduced a sheaf theoretic model and a quantitative measure that combined the syntactic structure of language with the probability distribution of its statistical patterns. When applied to garden path sentences and compared to surprisal, our model correlated better with human behavioural data, provided better predictions of human reading times and distinguished between different types of sentences.

Our results were, however, slightly underestimating: the human garden path effect was 313 ms and the model average 45.5 ms. We conjecture this is due to the presence of other linguistic features such as semantics and pragmatics. These features have been controlled for in other Psycholinguistic datasets and experimental results are also available for them, for instance see (Pickering and Traxler, 1998). There also exists related work on using machine learning to model them. For instance, the work of (Padó et al., 2009) implements a clustering algorithm to model plausibility. Improving on these results by using LLMs and presheaves is work in progress. We are also aiming to generalise the results of this paper by working on a large scale garden path dataset, developed recently in (Huang et al., 2023).

It is tempting to conclude that **IF** is a better estimator of human reading times for garden path data only, whereas surprisal works better for naturalistic data. However, we believe that this is not the case. Indeed, garden path sentences are also naturally occurring; they are the focus of our study since they reveal some essential properties of human sentence processing (also see (Huang et al., 2023) on this). The properties we work with in this paper are the human use of syntactic as well as lexical prediction when processing and understanding sentences. Our results show that whereas LLMS only use lexical predication, it is possible to endow with structures that can also model syntax. The resulting model thus uses both of these for prediction; consequently, it correlates better with human behavioural data.

The compositional nature of sheaves allows us to incorporate other features in the model, via pairing or the composition of their corresponding map. Less is known when it comes to the combination of their distance measures. Extending the model and further evaluations constitutes work in progress.

References

Samson Abramsky and Adam Brandenburger. 2011. The sheaf-theoretic structure of non-locality and contextuality. *New J. Phys.*, 13:113036.

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)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rui Soares Barbosa. 2014. On monogamy of non-locality and macroscopic averages: examples and preliminary results. *Electronic Proceedings in Theoretical Computer Science*, 172:36–55.

Thomas Bever. 1970. The cognitive basis for linguistic structures. *Cognition and the Development of Language*, pages 279–362.

Cristian Bodnar, Francesco Di Giovanni, Benjamin Chamberlain, Pietro Lio, and Michael Bronstein.

2022. Neural sheaf diffusion: a topological perspective on heterophily and oversmoothing in GNNs. In *Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems*, volume 35.

Tai-Danae Bradley, John Terilla, and Yiannis Vlassopoulos. 2022. An enriched category theory of language: From syntax to semantics. *La Matematica*, 1:551—-580.

Fernanda Ferreira and John M. Henderson. 1990. Use of verb information in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4):725–745.

Lyn Frazier. 1979. *On comprehending sentences: Syntactic parsing strategies*. Ph.D. thesis, Doctoral dissertation, University of Connecticut.

Lyn Frazier. 1987a. Sentence processing: A tutorial review. In Max Coltheart, editor, *Attention and Performance*, volume 12, pages 559–586. Erlbaum, Hillsdale, NJ.

Lyn Frazier. 1987b. *Sentence processing: A tutorial review.*, pages 559–586. Attention and performance 12: The psychology of reading. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US.

Lyn Frazier and Keith Rayner. 1990. Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, 29.

Susan M. Garnsey, Neal J. Pearlmutter, Emily Myers, and Melanie A. Lotocky. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37:58–93.

Alexander Grothendieck. 1957. Sur quelques points d'algèbre homologique, I. *Tohoku Mathematical Journal*, 9(2):119 – 221.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, page 1–8, USA. Association for Computational Linguistics.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2023. Surprisal does not explain syntactic disambiguation difficulty: evidence from a large-scale benchmark.

Steve Huntsman, Michael Robinson, and Ludmilla Huntsman. 2024. Prospects for inconsistency detection using large language models and sheaves.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

S.M. Lane, S. MacLane, and I. Moerdijk. 1992. *Sheaves in Geometry and Logic: A First Introduction to Topos Theory*. Hochschultext / Universitext. Springer-Verlag.

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

Kin Ian Lo, Mehrnoosh Sadrzadeh, and Shane Mansfield. 2022. A model of anaphoric ambiguities using sheaf theoretic quantum-like contextuality and bert. In *Proceedings End-to-End Compositional Models of Vector-Based Semantics,* NUI Galway, 15-16 August 2022, volume 366 of *Electronic Proceedings in Theoretical Computer Science*, pages 23–34. Open Publishing Association.

Ulrike Padó, Matthew W. Crocker, and Frank Keller. 2009. A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, 33(5):794–838.

Steven Philips. 2019. A universal construction for semantic compositionality. *Phil. Trans. R. Soc.*, B375.

Martin Pickering and Steven Frisson. 2001. Processing Ambiguous Verbs: Evidence from Eye Movements. *Journal of experimental psychology. Learning, memory, and cognition*, 27:556–73.

Martin J Pickering and Matthew J Traxler. 1998. Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(4):940.

Grusha Prasad and Tal Linzen. 2021. Rapid syntactic adaptation in self-paced reading: Detectable, but only with many participants. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(7):1156.

Bradley L. Pritchett. 1988. Garden path phenomena and the grammatical basis of language processing. *Language*, 64(3):539–576.

Michael Robinson. 2017. Sheaves are the canonical data structure for sensor integration. *Information Fusion*, 36:208–224.

M. Van Schijndel and T. Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *CogSci*.

Patrick Sturt, Martin J Pickering, and Matthew W Crocker. 1999. Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, 40(1):136–150.

Marten van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988.

D. Wang, M. Sadrzadeh, S. Abramsky, and V. Cervantes. 2021a. Analysing Ambiguous Nouns and Verbs with Quantum Contextuality Tools. *Journal of Cognitive Science*, 22(3):391–420.

D. Wang, M. Sadrzadeh, S. Abramsky, and V. Cervantes. 2021b. On the Quantum-like Contextuality of Ambiguous Phrases. In *Proceedings of the 2021 Workshop on Semantic Spaces at the Intersection of NLP, Physics, and Cognitive Science*, page 42–52. Association for Computational Linguistics.

Daphne Wang, Mehrnoosh Sadrzadeh, Miloš Stanojević, Wing-Yee Chow, and Richard Breheny. 2024. How can large language models become more human? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 166–176, Bangkok, Thailand. Association for Computational Linguistics.

Daphne Pauline Wang. 2024. *A Quantum-Inspired Analysis of Human Disambiguation Processes: Foundational Theory and Applications*. Ph.D. thesis, UCL (University College London).