How can large language models become more human?

Daphne Wang
University College London
Quandela

Mehrnoosh Sadrzadeh
University College London

Miloš Stanojević
University College London

Wing-Yee Chow
University College London

Richard Breheny
University College London

Abstract

Psycholinguistic experiments reveal that efficiency of human language use is founded on predictions at both syntactic and lexical levels. Previous models of human prediction exploiting LLMs have used an information theoretic measure called *surprisal*, with success on naturalistic text in a wide variety of languages, but under-performance on challenging text such as garden path sentences. This paper introduces a novel framework that combines the lexical predictions of an LLM with the syntactic structures provided by a dependency parser. The framework gives rise to an *Incompatibility Fraction*. When tested on two garden path datasets, it correlated well with human reading times, distinguished between easy and hard garden path, and outperformed surprisal.

1 Introduction

Psycholinguistic research develops models of human language understanding using experimental techniques such as self-paced reading and eye-tracking. Natural Language Processing research develops algorithms that enable machines solve human language tasks. Novel lines of research bringing these two fields together have emerged, where a question of interest has been whether machines are able to process language in ways similar to humans. The goal of this paper is to show that the answer can be yes, but only when they are equipped with human capabilities that enable them to predict with a combination of both syntactic structure and lexical statistics.

In order to model these characteristics, one needs a computational framework with at least two levels (more if we take pragmatics and other language features into account). We work with *presheaves* and specific instances of them, which consist of (1) a base that models linear structure, and (2) data that encode the statistics of different interpretations of the base. The data can be manifold recording outcomes of events, which can themselves be binary or many-valued, and their probabilities. For these reasons, presheaves provide a good candidate framework for modelling features of human language understanding.

We use a simple topological space as the base of our presheaf: that of a pre-ordered set. The elements of this set are sub-phrases of a sentence. The pre-order relation over the elements is the prefix relation between the sub-phrases. This relation will be used to the represent the incrementality of the parsing process. Our data is the probabilities of syntactic structures of sub-phrases. First, we obtained completions and their statistical information from the predictions of the large language model `GPT-2`. Then, to get the syntactic structures of the sub-phrases, we use the dependency parser `spaCy`. Our sheaf theoretic framework gives rise to a schematic fraction that measures how incompatible is the syntactic probability of a phrase from its completions. We refer to this fraction as the *incompatibility fraction* (**IF**). Well known distance measures between probability distributions exist and can be used when instantiating **IF**; we worked with `Kullback-Leibler divergence` (KL), `Jensen-Shannon divergence` (JS), and a measure similar to `Earth Movers` (EM).

Deep learning algorithms, especially attention-based ones, have made impressive advances in predicting the next words of a sentence. A statistical quantity known as "surprisal" has been found to correlate with human reading times (Levy, 2008; Hale, 2003, 2006). This, however, has only been the case for naturalistic text such as news paper articles. The jury is still out regarding a class of challenging sentences known as garden path (GP) sentences (Bever, 1970; Frazier, 1987; Frazier and Rayner, 1982). Psycholinguistic research has shown that humans experience processing difficulty and show longer reading times when processing GP sentences. Further, different types of

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 166–176
August 15, 2024 ©2024 Association for Computational Linguistics

syntactic ambiguities have been shown to result in different levels of processing difficulty (Sturt et al., 1999). So far, surprisal has not been able to accurately predict the human reading times of GP sentences and more importantly has not been able to distinguish between easy versus hard sentences (Schijndel and Linzen, 2018; van Schijndel and Linzen, 2021; Huang et al., 2023).

In order to test the applicability of our framework, we tested it on two GP datasets (Pickering and Traxler, 1998), with hard (i.e., subordinate clause) and easy (i.e., complement clause) ambiguities. Both datasets had a disambiguated control for each of their GP sentences. They also had variants of them which were either semantically plausible or implausible. **IF** was measured for all these sentences and its predictions compared with with human reading times and surprisal. All of the instances we worked with, i.e. KL, JS, and EM, correlated well with human reading times and had very low errors, predicted the differences between GP sentences and their disambiguated controls well, could distinguish between easy and hard garden path, and outperformed surprisal. On the semantic front, all the measures including surprisal validated one of the hypotheses, that a semantically implausible sub-phrase take longer to read. The other hypothesis was about shorter GP effects in implausible sentences, which could not be detected by any of the measures. Dealing with these needs an explicit encoding of the semantic structure of sentences and we believe presheaves can also help. Working out the details is left to future work.

2 Related Work

Inspired by applications of information theory to Psycholinguistics (Attneave, 1959), Hale argued that *suprisal* is a good measure for the cognitive load faced by humans during sentence processing (Hale, 2001, 2003, 2006). Surprisal measures the degree of unpredictability of a word w given its prefix context $w_1 \cdots w_n$ and is computed via the following formula:

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

Hale argued in favour of the use of surprisal in incremental parsing procedures. Building on this, Levy (2008) and later Smith and Levy (2013) showed that surprisal can also model the cognitive load modelled by constraint-based theories. The focus of Hale's work was on GP sentences, but he only provided experimental data for a couple of examples. Large scale validations on large datasets (Levy, 2008; Smith and Levy, 2013) and eleven different languages from five different language families followed suit (Wilcox et al., 2023). These only considered naturalistic text such as Wikipedia and news articles. Large scale data for GP sentences were not taken into account until more recent times (Schijndel and Linzen, 2018; van Schijndel and Linzen, 2021; Huang et al., 2023), where it was found out that surprisal does not provide good correlation. This has been the case for the surprisal computed over either syntactic predictions of a probabilistic parser or the lexical predictions of a statistical language model. In either case, the predictions largely underestimated human reading times. Weighted combinations of the syntactic and lexical surprisal were also computed but still underestimated (Arehalli et al., 2022). Another drawback of surprisal is that it has been unable to distinguish between easy and hard GP sentences.

Much of the original work on GP sentences focused on structural ambiguities. Here we have the original work, insights and examples of Bever (Bever, 1970), which was followed by the indepth analysis of Frazier (Frazier, 1979, 1987; Frazier and Rayner, 1990). Later work brought the role of semantics into the forefront. Since humans process language incrementally, it was expected that the existence of relevant semantic information would increase the speed of recovery from a local ambiguity. In this regard, Altmann et al. (1992); Altmann and Steedman (1988) studied the role of referential information, Trueswell et al. (1994) worked on the tenses of the verbs, and Pickering and Traxler (1998) on the lexical information encoded in sentential sub-phrases such as subject-verb and verb-object. Most of this work has only been verified by Psycholinguistic experiments on human subjects, but some of it was also verified using statistical machine learning methods such as clustering (Padó et al., 2009).

Presheaves and sheaves are general mathematical models introduced to formalise and reason about abstract notions of global consistency. They originate from the work of Jean Leray (Leray, 1959), whose aim was to study partial differential equations from a purely topological perspective. Subsequent work then extended the use of sheaf theory to other areas of mathematics, such as algebraic geometry (Cartan, 1950; Serre, 1955; Grothendieck, 1957) and logic (Lawvere, 1970; Tierney, 2011). More recently, sheaves and presheaves have been applied to formalise the consistency of different forms of concrete data. Here we have examples of data coming from quantum mechanics (Abramsky and Brandenburger, 2011), signal processing (Robinson, 2017), graph neural networks (Bodnar et al., 2022), and natural language (Wang et al., 2021a,b; Lo et al., 2022; Huntsman et al., 2024; Philips, 2019; Bradley et al., 2022). Notably, measures similar to **IF** were developed for physical experiments to compute the amount of unsharpness of experimental data (Vallée et al., 2024). These were preliminarily also tested on linguistic data, e.g. for the interpretations of phrases with semantic and anaphoric ambiguities (Wang et al., 2021a; Lo et al., 2022, 2023). A recent paper explores their applicability to ambiguities arising in garden path sentences but does not consider the general case nor the range of instantiations we offer here, works with the masked feature of BERT and has not been tested on semantic plausibility (Wang and Sadrzadeh, 2024).

3 Methodology

We use topological spaces and their associated data to model the sub-phrases of a sentence and their interpretations. The topological spaces model the relation between the sub-phrases as they are read by a human subject from a piece of text, i.e. incrementally and according to the linear flow of time. This order is also known as the *prefix order* or the *information order*. The data associated to each sub-phrase models the possible different interpretations of each sub-phrase and their probabilities. Here, we work with the completions of sub-phrases into a sentence and the probability of their syntactic structures. This is obtained via a combination of `GPT-2` and `spaCy` (with transformers). In what follows, we first go over the abstract model, then instantiate it to the concrete data of natural language, finally develop a set of measure that compute the differences between the different interpretations, giving rise to the notion of an **Incompatibility Fraction**.

3.1 Abstract Model

A topological space $\mathcal{X}$ is a tuple (X, τ) where X is a set of *points* and $\tau \subset \mathcal{P}(X)$ is the set of *open sets* which contains the empty set and is closed under arbitrary unions and finite intersections.

The open sets of a topological space can also have data associated to them. These are formalised through the notion of a *presheaf*, which is a map P that sends each subset U of X to the set PU of its data. The elements of the set PU are called *sections* over U, and can be seen as the possible data points on U. Here, we are interested in *events* and the *event presheaf* defined as follows. Given a set O of outputs (e.g. syntactic or semantic structures), an event is a map of the type $s \colon U \to O$. Whenever V is a subset of U, i.e. $V \subseteq U$, the presheaf *restricts* PU, i.e. the data points on U, to PV, i.e. the data points on V. For each element of $s \in PU$, the restriction is denoted by $s|_V$. This procedure is depicted in Fig. 1.

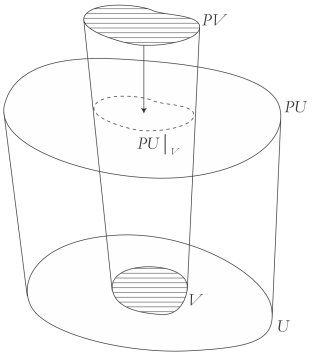


Figure 1: The restriction map of a presheaf.

Presheaves define a notion of *consistency* within sets via restriction maps. Consistency can also be defined across different sets. Given a presheaf P over a topological space $\mathcal{X}$, we say that there is *a gluing* between two sections $s_U \in PU$ and $s_V \in PV$ iff s_U and s_V are *locally consistent* or *compatible*, i.e. $s_U|_{U\cap V} = s_V|_{U\cap V}$. This definition leads to the fact that if there exists a gluing between two sections in PU and PV, then there will be an intersection between their restrictions $PU|_{U\cap V}$ and $PV|_{U\cap V}$, see Fig.2.

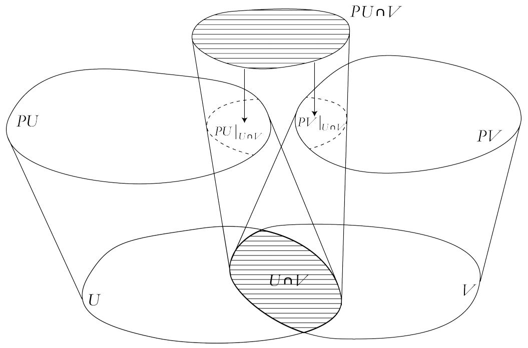


Figure 2: The presheaf structure over intersecting sets.

In order to model probabilistic events, an event presheaf P is post-composed with a distribution map $\mathcal{D}$ giving rise to a probabilistic event presheaf $\mathcal{D}P$. To a subset U of $\mathcal{X}$ the probabilistic presheaf assigns a set of probability distributions $\{d \mid d : U \to \mathbf{R}^+\}$. Whenever $V \subseteq U$, it computes the marginals of the probabilities of elements of U when restricted to V. Formally, this is as follows:

$$d_V(v) = \sum_{u \in V} d_U(u)$$

These probabilities are measured over our original set of outcomes O, via the principles events of the framework, i.e. $s \colon U \to O$.

3.2 Concrete Model

In the context of human sentence processing, our topological space $\mathcal{X}$ is the set of all incremental sub-phrases of the sentence under consideration. The order of the topology is the prefix relation over the sub-phrases of this sentence. Formally speaking, given the vocabulary σ of the sentences and σ^* the set of phrases over it, for $a, b, c, \dots \in \sigma^*$, we have

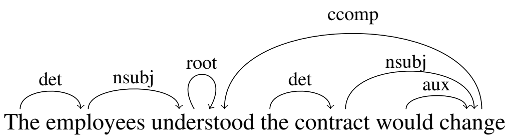
$$a \leq ab \leq abc \leq \cdots$$

As an example consider the sentence "The employees understood the contract", where we have the following instances of the prefix ordering:

The employees $\leq$ *The employees understood* $\leq$ *The employees understood the contract* $\leq$ *The employees understood the contract would change*.

In this sentence, however, there is no order relation between sub-phrases such as "The employees" and "employees understood". Despite the fact that they share "employees", none of them is a prefix of the other.

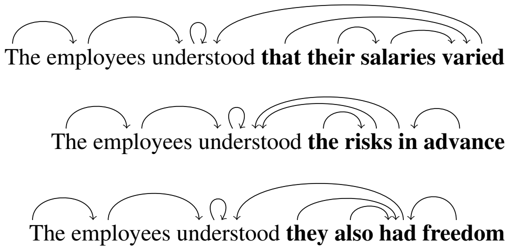
For the purposes of the current paper, we focus on a *syntactic* event presheaf, which assigns syntactic structures to completions of the sub-phrases into a full sentence. A section of the probabilistic event presheaf $\mathcal{D}P$ will then consist of a probability distribution over the syntactic structures of these completions. The syntactic structures are obtained using the transformer version of the dependency parser `spaCy` (Choi et al., 2015; Robinson, 1970). This parser returns a single parse for a full sentence. For example, the dependency parse for the sentence "The employees understood the contract would change" is as follows:



The employees understood the contract would change

Figure 3: Dependency relations in the sentence *The employees understood the contract would change..*

The completions of the sub-phrases and their statistics are obtained using the `GPT-2` model. See below for three different completions of the sub-phrase of "The employee understood" and their dependency structures.

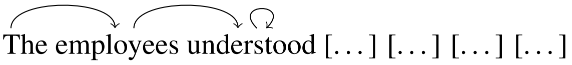


The employees understood **that their salaries varied**

The employees understood **the risks in advance**

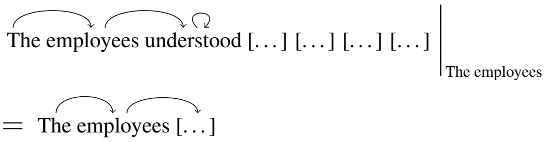
The employees understood **they also had freedom**

All of these lead to the same partial parse when restricted to the context "The employees understood", namely:



The employees understood [. . .] [. . .] [. . .] [. . .]

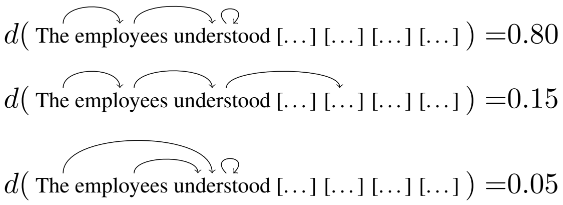
To obtain a syntactic structure for a sub-phrase, we use the restriction operations from the the presheaf, where we only keep the dependency information of each sub-phrase and ignore the rest of the sentence. For instance, the structure of the sub-phrase "The employees understood" restricted to "The employees" is obtained as follows:



The employees understood [. . .] [. . .] [. . .] [. . .] $\Big|_{\text{The employees}}$

= The employees [. . .]

The probability distributions associated to each parse are obtained from the predictions of `GPT-2` after sampling from 1000 instances and normalising the results. An example distribution is as follows:



$d($ The employees understood [. . .] [. . .] [. . .] [. . .] $)$ =0.80

$d($ The employees understood [. . .] [. . .] [. . .] [. . .] $)$ =0.15

$d($ The employees understood [. . .] [. . .] [. . .] [. . .] $)$ =0.05

Given two sub-phrases m_1 and m_1m_2 of $\mathcal{X}$ with $m_1 \leq m_1m_2$, suppose $d_{m_1m_2}$ is the probability distribution of the syntactic structures of m_1m_2. Then the restriction of $d_{m_1m_2}$ to m_1 for any syntactic structure $o \in O$ of m_1 is computed as follows:

$$d_{m_1m_2}|_{m_1}(o) = \Sigma_{o' \in O}\, d_{m_1m_2}(oo')$$

This restriction sums the probabilities of all completions of m_1 into m_1m_2, where m_1 retained the same syntactic structure after being completed by m_2. Note that, in general:

$$d_{m_1m_2}|_{m_1} \neq d_{m_1}$$

This is because the reader may have to do some reanalysis when going from m_1 to m_1m_2.

3.3 Measures

Each stage of the human reading process is modelled by a pair of succeeding sub-phrases of a sentence, e.g. (m_1, m_1m_2). The overall process of reading a sentence is modelled by a sequence of these pairs, i.e. $\{(m_i, m_{i+1})_j\}_{a \leq j \leq n-1}$ where n is the number of words or regions in a sentence. As an example, here is the first two pairs of a sequence that models the employee sentence:

(*The*, *The employees*)
(*The employees*, *The employees understood*)

As humans read an incoming sub-phrase m_1 of a sentence, they construct interpretations for it and assign probabilities to their interpretations. When the next region m_2 is read, a new set of interpretations and probabilities are constructed, this time for the sub-phrase m_1m_2. The reader expects that the interpretations and probabilities of m_1m_2 to be consistent with those of m_1. If this is the case, the sub-phrase m_1m_2 is comprehended and sentence processing can carry on linearly. For critical regions of GP sentences, however, this is not the case and as a result sentence processing is halted. This leads to a pause and possibly a reversal of the order of reading thus higher reading times are observed. Take our employee sentence and the pair of sub-phrases therein ("The employees understood the contract, The employees understood the contract would change"). This pair sits at the critical region of the garden path effect of the sentence. The shared prefix "The employees understood the contract" has a subject-verb-object structure in the first sub-phrase, which is not consistent with the subject-verb-subject structure after seeing "would change" in the second sub-phrase.

In order to check whether the structure and probabilities of the two succeeding sub-phrases m_1 and m_1m_2 of a sentence match, the larger sub-phrase m_1m_2 is restricted to the smaller one m_1 and the degree of their divergence is estimated. This divergence is what we refer to as the *Incompatibility Fraction* **IF**.

A common choice for measuring divergence is the Kullback–Leibler or KL-divergence. In our case, we measure the KL-divergence between a distribution d_{m_1} to $d_{m_1m_2|_{m_1}}$, given below:

$$KL(d_{m_1}||d_{m_1m_2|_{m_1}}) = \sum_o d_{m_1}(o) \log \frac{d_{m_1}(o)}{d_{m_1m_2|_{m_1}}(o)}$$

KL is not always defined, in which case its symmetric variant Jensen-Shannon divergence is used. In the interest of space will not provide the formula.

Another choice is a metric similar to what is known as Earth-Mover's and measures the overlap between two distributions by taking their min, i.e. $\sum_o \min(d_{m_1m_2}|_{m_1}(o), d_{m_1}(o))$. The divergence between the two distributions is then computed by subtracting the overlap from 1. This leaves us with the following formula:

$$1 - \sum_o \min(d_{m_1m_2}|_{m_1}(o), d_{m_1}(o))$$

All three of these instantiations can be used, giving rise to the following three measures:

$$\begin{array}{rc} \textbf{IF-min}: & 1 - \Sigma_o \min(d_{m_1}(o), d_{m_1m_2}|_{m_1}(o)) \\ \textbf{IF-KL}: & \mathrm{KL}(d_{m_1}||d_{m_1m_2}|_{m_1}) \\ \textbf{IF-JS}: & \mathrm{JS}(d_{m_1}||d_{m_1m_2}|_{m_1}) \end{array}$$

4 Experiments

We worked with two datasets put forwards by Pickering and Traxler in Pickering and Traxler (1998). Dataset 1 has GP sentences with complement clause ambiguities. An example is the following:

> Dataset 1. (i) GP. The dog catcher worried the terrier which fell wouldn't fit into the box.

Dataset 2 has GP sentences with subordinate-clause ambiguities. An example is the following:

	Equation	ρ	p-value
IF-min First Pass	$0.0018 \times \mathbf{IF}_{\text{min}} - 0.0776$	**0.595**	0.00032
IF-min Total	$0.0006 \times \mathbf{IF}_{\text{min}} + 0.14387$	0.448	0.00999
IF-JS First Pass	$0.0016 \times \mathbf{IF}_{JS} - 0.1333$	0.568	0.00068
IF-JS Total	$0.00053 \times \mathbf{IF}_{JS} + 0.0633$	0.4231	0.01580
IF-KL First Pass	$0.0066 \times \mathbf{IF}_{KL} - 0.4238$	0.445	0.0106
IF-KLTotal	$0.0021 \times \mathbf{IF}_{KL} + 0.4022$	0.326	0.06773
SP First Pass	$0.7361 \times \mathbf{SP} + 268.8467$	0.356	0.045
SP Total	$2.1326 \times \mathbf{SP} + 441.9445$	**0.459**	0.008

Table 1: Regression Equations with ρ's and their p-values.

Dataset 2. (i) GP. After the judge decided the verdict of the trial caught the old man's attention.

Each dataset has 24 sets of four sentences: (i) a plausible main sentence with a GP effect, and (ii) its disambiguated control, (iii) an implausible variant of the main sentence, and (iv) its disambiguated control. See below for examples of the disambiguated controls of Dataset 1. (i) GP and Dataset 2. (i) GP:

Dataset 1. (ii) DisAmb. The dog catcher worried that the terrier which fell wouldn't fit into the box.
Dataset 2. (ii) DisAmb. After the judge decided, the verdict of the trial caught the old man's attention.

The disambiguated controls of dataset 1 are obtained by adding a complementiser, such as 'that' to the garden path sentences. The disambiguated controls of dataset 2 are obtained by adding a comma. Sentences of dataset 1 are also known are as NP/S. They are an example of **easy** GP. Sentences of dataset 2 are known as NP/Z and are an example of **hard** GP.

The GP effect should occur after the second verb is encountered which we will refer to as the *critical region*, for example in "wouldn't fit in the box" in Dataset 1. (i) GP and in "caught the old man's attention" in Dataset 2. (i) GP.

Our hypothesis is that in either dataset, the reading times (both first-pass reading times and total reading times) of (i) sentences are longer than (ii) sentences. This is since the (ii) sentences are the disambiguated controls with no GP whereas the (i) sentences each contain a GP. A GP effect is computed by subtracting the reading time of (ii) sentences from the reading time of (i) sentences over the critical region. We expect that this effect is higher in Dataset 2 (which has hard GP sentences) than in Dataset 1 (which has easy GP sentences).

Items (iii) and (iv) differ from (i) and (ii) according to the plausibility of the sub-phrases preceding their critical regions. Here are examples of the implausible variants of the sentences from both datasets with their disambiguated controls:

Dataset 1. (iii) GP. The dog catcher worried the book which fell wouldn't fit into the box.
Dataset 1. (iv) DisAmb. The dog catcher worried that the book which fell wouldn't fit into the box.

Dataset 2. (iii) GP. After the judge packed the verdict of the trial caught the old man's attention.
Dataset 2. (iv) DisAmb. After the judge packed, the verdict of the trial caught the old man's attention.

The difference in plausibility has an impact on the magnitude of the GP effect. Here, we have two hypotheses: first that the garden path effects of the these, e.g. (iii), in either Dataset 1 or 2, are shorter than the ones without them, e.g. (i), and second that, the total reading times of implausible sentences are longer when the implausibility occurs, e.g. in "the book which fell" in Dataset 1. (iii). GP or "the verdict of the trial" in Dataset 2. (iii) GP; we will refer to this region as the *plausibility region*. The reason for hypothesis 1 is that the implausibility is designed to diminish the misanalysis and lead to a smaller GP effects. Indeed, it was shown in Pickering and Traxler (1998) that GP sentence with implausible prefixes exhibit a smaller effect as compared to plausible ones, since the reader will be less inclined to "take the garden path". The

	All		**Hard** (NP/Z) **GP**		**Easy** (NP/S) **GP**	
Method	GPE	SE	GPE	SE	GPE	SE
IF-min First Pass	**39.47**	**0.17**	53.94	2.72	24.99	2.74
IF-min Total	66.17	10.92	90.44	11.18	41.90	10.65
IF-JS First Pass	39.69	0.43	**52.22**	**2.40**	**27.16**	**2.31**
IF-JS Total	65.73	10.94	86.49	11.35	44.97	10.51
IF-KL First Pass	52.81	3.64	62.20	4	43.42	3.30
IF-KL Total	86.28	9.96	**101.62**	**10.67**	**70.94**	**9.19**
Surprisal First Pass	0.35	0.16	0.72	0.32	-0.02	0.05
Surprisal Total	1.01	0.47	2.10	0.92	-0.07	0.16
Human First Pass	39.5		46.5		32.5	
Human Total	185.5		215.5		155.5	

Table 2: Garden Path Effects (GPE) and their Standard Errors (SE). All numbers are in milliseconds.

reason for hypothesis 2 is simply that implausible sentences are harder to comprehend than plausible ones, hence producing a slowdown in reading times when the implausibility is encountered. However, it was shown in Pickering and Traxler (1998) that this slowdown is more marked in the total reading times, and less effect is found in the first-pass.

5 Results and Analysis

We trained a regression model between human first pass and total reading times for all of the regions in all sentences and each of our distance measures. The individual regression equations, their resulting degrees of correlations and corresponding p-values are presented in in Table 1. All of our **IF** measures achieved a high correlation with both human reading times. In most cases these correlations were statistically significant. **IF**-min provided the highest and most significant correlations for first-pass reading, closely followed first by **IF**-JS, **IF**-KL and then surprisal. On the other hand, surprisal appears to correlate better with total reading time, although both the correlation coefficient and p-values are comparable with the ones obtained for **IF**-min. This means that **IF**-min is a good predictor of human reading times, and more specifically that they are better predictors of first-pass reading-times.

The individual regression models were used to predict reading times for sentences of types (i)-(iv). Given that our **IF** measures are all well correlated with human reading times, we expect to observe a significant difference between the ambiguous and unambiguous sentences, i.e. a high garden path effect (GPE). This is presented in column “All” of Table 2. **IF**-min achieves the best results with a high GPE of 39.47 millisecond and the lowest standard errors (SE) of 0.17. Although surprisal correlated well with reading times in general, it predicted very low GPE’s, and sometimes does not even predict the existence of a garden path-effect (notably for NP/S sentences). This shows that our measures indeed perform better than surprisal in predicting the garden path effects.

The GPE of hard versus easy sentences are presented in columns “NP/Z” and “NP/S” of Table 2, respectively. We expect to see a higher GPE for hard sentences. This is indeed the case for all models. The GPE’s of NP/Z column are higher than the GPE’s of NP/S columns. Our best measure for this distinction were **IF**-JS for first pass reading times and IF-KL for total reading times. They both predicted their GPE’s with the lowest error. All of the **IF** measures outperformed surprisal, which had the highest errors with the overall GPE. This was also individually the case for each of our tests: (1) our NP/Z test had an SE of 6.79 for first pass and an SE of 14.54 for total reading times, (2) our NP/S test had an SE of 5.68 for first pass and an SE of 12.39 for total reading times. Overall, all the models predicted the first pass reading times better than the total ones.

So far we have only considered syntactic effects. In order to evaluate whether our model is able to detect some semantic effects, we study the predictions for plausible and implausible sentences. The reading times for plausible and implausible sentences are in Tables 3 and 4. The results in Table 3 show that none of the measures could predict that GPE’s diminish with implausible cues. In fact, all of the measures showed the opposite. As we can see, the GPE's of implausible sentences are all higher than for plausible ones.

	Plausible	Implausible
Method	GPE	GPE
IF-min First Pass	28.66	50.28
IF-min Total	48.05	84.29
IF-JS First Pass	26.55	52.83
IF-JS Total	43.98	87.48
IF-KL First Pass	39.58	66.05
IF-KL Total	64.66	107.90
Surprisal First Pass	-0.11	0.81
Surprisal Total	-0.33	2.35
Human First Pass	50.5	28.5
Human Total	265	106

Table 3: Garden Path Effects (GPE) for plausible and implausible sentences. All numbers are in milliseconds.

	Plausible	Implausible
Method	RT	RT
IF-min First Pass	565.03	597.69
IF-min Total	988.30	1043.06
IF-JS First Pass	562.63	589.18
IF-JS Total	984.75	1028.75
IF-KL First Pass	560.50	578.13
IF-KL Total	981.79	1010.60
Surprisal First Pass	616.43	616.53
Surprisal Total	1112.01	1112.30
Human First Pass	673.5	686.25
Human Total	1222.5	1275.75

Table 4: Reading Time (RT) for plausible and implausible sentences (over the plausiblity region). All numbers are in milliseconds.

Table 4 shows that all of the **IF** measures could however verify our second hypothesis. As we can see, all the measures, including surprisal, predicted a longer reading time for implausible sub-phrases, although the differences where much more marked in the case of the **IF** measures. Indeed, although the absolute values of the surprisal predictions are closer to the human baseline, the differences in predicted reading times of plausible and implausible sentences were closer to the observed human one for using the **IF** measures. For the first pass reading times, this difference in the human time was 12.75 ms. All the **IF** measures predicted a similar distance; the lowest predicted difference was KL with a difference of 17.63 ms and the higher used **IF**-min with a difference of 32.66 ms. Surprisal, on the other hand, predicted a very low difference of 0.10 ms. Regarding the total reading times: the difference in human times was 53.25 ms; **IF**-min was our best measure, which predicted a difference of 54.76 ms, followed by KL with a prediction of 44 ms and finally JS with 28.81 ms. Surprisal came last, with a very low difference of 0.29 ms.

6 Conclusions and Future Work

Our work highlights the importance of combining syntactic structure and lexical statistics when modelling human language understanding. For syntactic structure we worked with the linear prefix ordering between sub-phrases of a sentence and their dependency structures. For lexical statistics, we worked with sub-phrase completions and their probabilities provided by an LLM. An incompatibility fraction was developed to measure the distance between probability distributions of sub-phrases and their completions. We experimented with known relative entropy distances (KL and JS) and Earth Movers, all of which showed a strong correlation with human behaviour in syntactic GP sentences and outperformed surprisal. None of the measures however, neither any of ours nor surprisal, were successful when it came to GP sentences with semantic implausibilities. We believe these sentences are too complex and in order to deal with them, one needs to explicitly model semantic structure. As it is, the predictions of the parser are over shadowed by the probabilities provided by the LLM, which predicts very high incompatibility and surprisal for implausible phrases.

Kullback–Leibler has a long history of applications in natural language tasks, e.g. in measuring the semantic content of words (Herbelot and Ganesalingam, 2013) and deriving objective functions for language models(Labeau and Cohen, 2019). Notably, Levy showed that under certain assumptions it equates surprisal (Levy, 2008). `Earth Movers` has also been applied in Natural Language Processing, e.g. to compute the relationship between a document and its words (Kusner et al., 2015) and the distance between bilingual lexicons (Huang et al., 2016; Zhang et al., 2016). The main difference between the modelling part of these works and ours is the measurement events. We work with sub-phrases and their syntactic structures, whereas the other measures only consider word co-occurrence. Despite these, we believe there should be a relationship between the incompatibility of two phrases and their degree of surprisal. Formalising this relation is work in progress.

There are four other directions that we aim to pursue in future work. These are as follows: (I) The focus of the paper was on garden path sentences. More work is required to test the performance of our measures on a wider range of naturally occurring sentences. (II) The plausibility element of the dataset used in this work may not be representative of the garden-path effect as a whole. We therefore also plan to replicate our results using different datasets, notably the ones of (Huang et al., 2023; Prasad and Linzen, 2021) (III) As structure, we only considered syntax. Modelling semantic structure of sub-phrases and sentences, e.g. as agent-patient relations or event structures and/or the thematic information associated with verbs needs to be done. (IV) Our framework is by default only forward looking; experimenting with regression and back tracking to model repair and recovery is left to future work.

References

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

Gerry Altmann and Mark Steedman. 1988. Interaction with context during human sentence processing. *Cognition*, 30(3):191–238.

Gerry T.M Altmann, Alan Garnham, and Yvette Dennis. 1992. Avoiding the garden path: Eye movements in context. *Journal of Memory and Language*, 31(5):685–712.

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.

Fred Attneave. 1959. *Applications of Information Theory to Psychology: A summary of basic concepts, methods and results*. Holt, Rinehart and Winston.

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.

Henri Cartan. 1950. Idéaux et modules de fonctions analytiques de variables complexes. *Bulletin de la Société mathématique de France*, 78:29–64.

Jinho D Choi, Joel Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 387–396.

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

Lyn Frazier. 1987. *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. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2):178–210.

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

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. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123.

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

Aurélie Herbelot and Mohan Ganesalingam. 2013. Measuring semantic content in distributional vectors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445, Sofia, Bulgaria. Association for Computational Linguistics.

Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. 2016. Supervised word mover's distance. In *Advances in Neural Information Processing Systems (NIPS)*.

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.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966. PMLR.

Matthieu Labeau and Shay B. Cohen. 2019. Experimenting with power divergences for language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4104–4114, Hong Kong, China. Association for Computational Linguistics.

F William Lawvere. 1970. Quantifiers and sheaves. In *Actes du congres international des mathematiciens, Nice*, volume 1, pages 329–334.

Jean Leray. 1959. Théorie des points fixes : indice total et nombre de lefschetz. *Bulletin de la Société Mathématique de France*, 87:221–233.

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.

Kin Ian Lo, Mehrnoosh Sadrzadeh, and Shane Mansfield. 2023. Generalised winograd schema and its contextuality. In *Proceedings of 20th International Conference on Quantum Physics and Logic*, Institut Henri Poincare, Paris, France.

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 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.

Jane J. Robinson. 1970. Dependency structures and transformational rules. *Language*, 46(2):259–285.

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*.

Jean-Pierre Serre. 1955. Faisceaux algébriques cohérents. *Annals of Mathematics*, pages 197–278.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

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.

M. Tierney. 2011. *Axiomatic Sheaf Theory : Some Constructions and Applications*, pages 249–326. Springer Berlin Heidelberg, Berlin, Heidelberg.

John C. Trueswell, Michael K. Tanenhaus, and Susan M. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33(3):285–318.

Kim Vallée, Pierre-Emmanuel Emeriau, Boris Bourdoncle, Adel Sohbi, Shane Mansfield, and Damian Markham. 2024. Corrected bell and non-contextuality inequalities for realistic experiments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382(2268):20230011.

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 and Mehrnoosh Sadrzadeh. 2024. Causality and signalling of garden-path sentences. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382(2268):20230013.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Meng Zhang, Yang Liu, Huanbo Luan, Maosong Sun, Tatsuya Izuha, and Jie Hao. 2016. Building earth mover's distance on bilingual word embeddings for machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.