PHILOSOPHICAL
TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

Cite this article: Wang D, Sadrzadeh M. 2024 Causality and signalling of garden-path sentences. *Phil. Trans. R. Soc. A* **382**: 20230013. https://doi.org/10.1098/rsta.2023.0013

Received: 29 May 2023
Accepted: 1 September 2023

One contribution of 12 to a theme issue 'Quantum contextuality, causality and freedom of choice'.

Subject Areas:
artificial intelligence, quantum computing

Keywords:
sheaf theory, causality, contextuality, natural language ambiguities, psycholinguistics, garden path phenomena

Author for correspondence:
Mehrnoosh Sadrzadeh
e-mail: m.sadrzadeh@ucl.ac.uk

THE ROYAL SOCIETY
PUBLISHING

Causality and signalling of garden-path sentences

Daphne Wang and Mehrnoosh Sadrzadeh

Department of Computer Science, University College London, London, UK

DW, 0000-0001-5088-2759; MS, 0000-0002-5863-7835

Sheaves are mathematical objects that describe the globally compatible data associated with open sets of a topological space. Original examples of sheaves were continuous functions; later they also became powerful tools in algebraic geometry, as well as logic and set theory. More recently, sheaves have been applied to the theory of contextuality in quantum mechanics. Whenever the local data are not necessarily compatible, sheaves are replaced by the simpler setting of presheaves. In previous work, we used presheaves to model lexically ambiguous phrases in natural language and identified the order of their disambiguation. In the work presented here, we model syntactic ambiguities and study a phenomenon in human parsing called garden-pathing. It has been shown that the information-theoretic quantity known as 'surprisal' correlates with human reading times in natural language but fails to do so in garden-path sentences. We compute the degree of signalling in our presheaves using probabilities from the large language model BERT and evaluate predictions on two psycholinguistic datasets. Our degree of signalling outperforms surprisal in two ways: (i) it distinguishes between hard and easy garden-path sentences (with a p-value $<10^{-5}$), whereas existing work could not, (ii) its garden-path effect is larger in one of the datasets (32 ms versus 8.75 ms per word), leading to better prediction accuracies.

This article is part of the theme issue 'Quantum contextuality, causality and freedom of choice'.

1. Introduction

Natural language ambiguities give rise to probability distributions and these can be studied in the mathematical framework of sheaf theory. Sheaf-theoretic models go all the way back to the work of Jean Leray during the second world war, and his formalization of fixed points of partial differential equations [1]. Later, it was Alexander Grothendieck that studied them in detail to unify different fields with mathematics [2]. In [3], Abramsky and Brandenburger showed that sheaf theory can also be applied to model the notion of contextuality in physics, and reason about paradoxes such as violations of Bell inequalities. In following works, Kishida *et al.* used sheaf theory to formalize a more general notion of paradox and in particular modelled famous logical paradoxes, such as the liar's paradox [4,5]. Our previous work expanded on the latter and further showed that sheaf theory can model the manifold of interpretations arising from lexical ambiguities in natural language [6–8].

Natural language ambiguities are not limited to lexical ambiguities. They also occur in syntactic analysis, i.e. when a phrase can have multiple syntactic structures, or in discourse, where a pronoun or ellipsis marker can have more than one referent. In our preliminary work on the sheaf models of discourse, we showed that the ambiguities in the Winograd schema challenge, a pinnacle challenge of Natural Language Processing (NLP), can be extended to scenarios that exhibit quantum-like contextuality [9,10]. In the current paper, we show that sheaf theory can also model an interesting phenomenon in psycholinguistics, known as 'the garden-path effect'. Garden-path sentences first appeared in [11] and sparked a research program on the biological and cognitive bases of human interaction patterns. The 1980s and 1990s saw a wave of psycholinguistic experiments aiming to classify these ambiguities and measure their cognitive dissonance levels. Recently, NLP researchers have found correlations between human reading times and statistics learnt by large language models [12]. However, these methods do not give as good results for garden-path sentences [13], and it remains an open problem to understand why that is the case.

Our choice of sheaf theory is motivated by its usage in quantum scenarios [3], which also motivated our previous study of lexical ambiguities [6–8]. The sheaf condition of quantum scenarios corresponds to the no-signalling condition in non-locality protocols. No-signalling can be imposed on physical systems, for instance by isolating the parties involved. This is much harder to impose on linguistic scenarios. In the absence of no-signalling, sheaf theory cannot straightforwardly be used. In order to remedy the situation, we previously used the setting of Contextuality-by-Default (CbD), which is able to analyse data coming from any scenario regardless of the no-signalling property [14,15]. We witnessed that linguistic data are able to show contextuality in the same ways quantum scenarios do (signalling property aside). More recently, the unsharpness of experimental apparatus inspired the sheaf theoretician to also factor in a *signalling fraction* in their models [9,10,16]. Working within this fraction, we were able to model discourse ambiguities and also find quantum contextual examples [9,10]. Another way of dealing with the presence of signalling is to study causal scenarios [17–19]. In previous work, we used causal models and the so-called *causal fraction* to further our study of lexical ambiguities [20]. In this paper, we focus on syntactic ambiguity and show that by building empirical models which imitate the incremental order of human sentence processing, the signalling fraction will correspond to the causal fraction associated with the reading order. We then make use of these fractions to predict human reading times of garden-path sentences and obtain improvement on existing research that uses an information theoretic measure known as surprisal.

2. Background

(a) Garden-path sentences

In 1970, the psycholinguist Thomas Bever introduced a class of natural language sentences which, although unambiguous as a whole, are difficult for humans to parse due to local ambiguities [11]. A preliminary study of these sentences revealed that the difficulty arises from human biases in language usage, coming from a set of factors such as frequency, plausibility and primordiality: these biases lead the human subject down a 'garden path' when reading these sentences. Which one of these is the main factor remains debated amongst researchers; for instance, Bever's view was that it is primordiality. The ambiguities studied by Bever were syntactic, in other words, they came from multiplicities of syntactic structures. Explicitly, a garden-path sentence is a sentence that has a single global syntactic structure but admits local syntactic ambiguities at some of its initial stages. In addition, a *garden-pathing* phenomenon is created in these sentences when the local structures which disambiguate the sentence have lower probabilities compared with the wrong ones. Examples of garden-path sentences are:

(1a) The employees understood the contract would change.
(1b) Because the employees negotiated the contract would change.

The locally ambiguous parts of these sentences are: 'The employees understood the contract' and 'Because the employees negotiated the contract'. The (English-speaking) human biases about these initial structures tell them that—most probably—the verbs 'understood' and 'negotiated' have 'the employees' as subjects, and 'the contract' as objects. These come from two possibilities (i) either subject–verb–object is the most frequent syntactic form of English sentences, (ii) or the relation actor–action–object is a primordial humane semantic relation. As a result, the subject–verb–object construction has been observed/used much more than any other construction and the reader is eager to enforce it to the first three components of every sentence. As the reader reads on, they realize that, in fact, this hasty decision is wrong as the next part of the sentence, i.e. 'would change', does not fit in. In the case of (1a), a reanalysis reveals that the main verb takes the sentential (S) complement 'the contract would change' as an object instead of the noun phrase (NP) 'the contract'. Such garden-path sentences, i.e. when an S is mistaken for an NP, are classified as NP/S. In the case of (1b), the reader eventually realizes that the main verb 'negotiated' has no (Zero) object, as opposed to the original analysis which designated the noun phrase (NP) 'the contract' as a complement. These sentences are classified as NP/Z. The above garden-path sentences are semantically equivalent to the following sentences, (2a) and (2b), which are much easier to understand. They will be referred to as the *unambiguous versions* of our garden-path sentences.

(2a) The employees understood *that* the contract would change.
(2b) Because the employees negotiated, the contract would change.

A major feature of garden-path sentences is a slowdown in reading time when entering the so-called *critical region* or *disambiguating region* of the sentence. In our examples, this is when reading the phrase 'would change'. No slowdown is observed in the same region of the unambiguous versions. This difference in reading times is referred to as the *garden-path effect*. Different classes of garden path sentences, i.e. NP/S and NP/Z, result in different garden-path effects; for instance, NP/Z sentences are significantly harder to comprehend and therefore have a larger garden-path effect, than NP/S sentences [21,22].

(b) Modelling garden-path with surprisal

Psycholinguistic studies have shown that one of the main factors influencing reading time is *predictability* of a word in context [23]. Words are read faster if found in a context that makes them predictable (e.g. 'shark' in the context 'The coast guard had warned that someone had seen a'), than in contexts where they are not (e.g. 'shark' in the context 'The zoo keeper explained that the lifespan of a'). In [12], it was shown that the relation between predictability and reading time is *logarithmic*. Formally speaking, *surprisal* is defined from the conditional probability of encountering a word w in the context $c = w_1 \dots w_n$ as:

$$\mathsf{SP}(w|w_1 \dots w_n) = -\log_2 P[w \mid w_1 \dots w_n].$$

Also known as self-information, surprisal originates in Shannon's theory of information [24], where it is defined as the *quantity of information* entailed by knowing the value $X = w$, where X is defined as a random variable selecting the next word in the context $w_1 \dots w_n$. The intuition is that a very predictable word is not surprising, and therefore does not carry out a lot of information; on the other hand, if a word is not predictable, it significantly increases the amount of information available to the reader. In psycholinguistics, surprisal is used as a predictor for reading time (RT), according to the following relation, pinned by [12]:

$$\mathrm{RT}(w|w_1 \dots w_n) \propto \mathsf{SP}(w|w_1 \dots w_n).$$

The above results were obtained by looking at eye-tracking times of a subset of the Dundee dataset, and self-paced reading times for subsets of the Brown corpus, both corpora containing naturalistic sentences. The idea of studying garden-path sentences using surprisal in fact predates these findings, and is attributed to the work of Hale [25]. In [25], he used the probabilities obtained from a probabilistic context-free parser to predict the existence of a garden-path effect. In the following work, empirical correlations between surprisal and self-paced reading times of garden-path sentences were also studied [13,26–28]. It was shown that, although the surprisal calculated using different language models is able to predict a garden-path effect, it consistently underestimates its magnitude. In addition, although predictions were mostly lower for NP/S than NP/Z [13,27], there was no statistical difference between the two; in fact, in [26], the average garden-path effect for NP/S sentences was lower than the one for NP/Z sentences.

(c) Sheaf theory and quantum contextuality and causality

Sheaf theory has been used to model and reason about a fundamental phenomenon in quantum mechanics known as *contextuality*. Contextuality was first introduced by Kochen & Specker [29] as an extension of the principle of non-locality [30,31]. In [3], it was shown that in the language of sheaves, contextuality corresponds to the impossibility of finding a global section compatible with a family of local sections. In this framework, the possible local measurements are taken from a set X, and a compatibility relation is imposed on this set. This compatibility relation essentially encodes which measurements can be simultaneously made; we note that, interpreted as such, this relation is symmetric, i.e. a can be performed at the same time as b if b can be performed in the same time as a. For example, in the standard (2,2,2)-Bell scenario (i.e. consisting of two parties, each choosing between two measurements, and each measurement having two possible outcomes), we start from a set of measurements $X = \{a_1, a_2, b_1, b_2\}$, where $I_A = \{a_i\}_{i=1,2}$ corresponds to the choice of measurements available to Alice, and $I_B = \{b_i\}_{i=1,2}$ is the set of measurements available to Bob. Then, any of Alice's measurements in I_A is compatible with any of Bob's, i.e. $b \in I_B$. However, the measurements a_1 and a_2 are not compatible, as they cannot be performed simultaneously (and similarly for Bob's measurements). Each of these measurements comes with a set of possible outcomes O.[1] Then, given a set of compatible measurements U, we can see an event as associating outcomes with the measurements selected in U. This can be seen as a function:

$$s : U \to O.$$

Formally speaking, these functions are modelled as the *presheaf of events*, i.e. a contravariant functor with the type $\mathcal{E} : \mathcal{C}^{op} \to \mathbf{Set}$, where $\mathcal{C}$ is, in general, a subcategory of $\mathcal{P}(X)$ with the inclusion relation. The action of this *presheaf* on objects U gives us the set of all possible *assignments* or *functions* $s : U \to O$. The action on morphisms $f : U \to V$ in $\mathcal{C}$ gives us all the *restrictions* of these

[1] Wlog, we can assume that O is the same for any choice of measurement.

context	BERT input
the	the [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
the employees	the employees [MASK] [MASK] [MASK] [MASK] [MASK]
the employees understood	the employees understood [MASK] [MASK] [MASK] [MASK]
the employees understood the	the employees understood the [MASK] [MASK] [MASK]
the employees understood the contract	the employees understood the contract [MASK] [MASK]
the employees understood the contract would	the employees understood the contract would [MASK]

Figure 1. BERT inputs for the sentence. *The employees understood the contract would change*.

assignments, namely:

$$\left.\begin{array}{ll} \mathrm{res}^V_U : \mathcal{E}(V) & \to \mathcal{E}(U) \\ s_V : V \to O & \mapsto s_V|_U : U \to O \\ v \mapsto o_v & u \mapsto o_{f(u)} \end{array}\right\} \qquad (2.1)$$

For example, in the (2,2,2)-Bell scenario, if Alice chooses to perform the measurement a_1 and obtains outcome $x \in O$, and Bob the measurement b_2 and obtains the outcome $y \in O$, then an event could be represented as the function:

$$s : U \to O :: a_1 \mapsto x; b_2 \mapsto y.$$

In quantum mechanics, however, the outcomes of measurements are not generally deterministic, so instead of looking at events, it is more relevant to look at *the probability distributions* over all of the possible events. This is obtained by post-composing the event presheaf $\mathcal{E}$ with the distribution monad $\mathcal{D}_{\mathbb{R}_+} : \mathbf{Sets} \to \mathbf{Sets}$, which gives back a presheaf and is defined as follows:

$$\begin{array}{rl} \text{on objects:} & \mathcal{D}_{\mathbb{R}_+} : \mathbf{Sets} \to \mathbf{Sets} :: U \mapsto \{d : U \mapsto U \to \mathbb{R}_+ \mid d \\ & \text{probability distribution over } U\} \\ \text{on morphisms:} & U \xrightarrow{f} V \mapsto \begin{array}{ccc} \mathcal{D}_{\mathbb{R}_+}(U) & \to & \mathcal{D}_{\mathbb{R}_+}(V) \\ d_U & \mapsto & d_V \text{ s.t. } d_V(v) = \sum_{u \in f^{-1}(v)} d_U(u). \end{array} \end{array}$$

Given an object $U \in \mathcal{C}$, an element of $\mathcal{D}_R\mathcal{E}(U)$ is called a *section* over U. This is where the notion of an *empirical model* comes in handy. These offer descriptions of the systems and each of them is the collection of probability distributions, over all of the possible events of a system. An example of an empirical model over the (2,2,2)-Bell scenario is depicted in figure 1. Formally speaking, given a set of objects $\mathcal{M}$ of $\mathcal{C}$, an *empirical model e* is a collection of sections of the presheaf $\mathcal{D}_R\mathcal{E}$ which selects a single probability distribution for each context $C \in \mathcal{M}$. The elements of $\mathcal{M}$ are the measurement contexts.

In the standard contextuality experiments, we are interested in studying the source of the correlations between contexts, i.e. choices of measurements, and their observed statistics. In order to isolate the source of potential correlations between the contexts and the outcomes, the standard practice is to limit the overall number of possible sources of such correlations. One type of correlation which can be eliminated in quantum experiments is communication, i.e. the *signalling* between Alice and Bob in the above example. In practice, this can be achieved by spatially isolating these parties. The consequence of such isolation, or lack of signalling, is that the marginal probability distributions do not depend on the choice of measurements of the other parties. In other words, for any set of inputs U, and any two sets of measurements V, V' compatible with all elements of U, we should have:

$$d_{U\cup V}|_U(\underline{o}_U) = d_{U\cup V'}|_U(\underline{o}_U) \qquad (2.2)$$

for all joint outcomes $\underline{o}$ over the measurements of U, where d_W corresponds to the joint probability distribution corresponding with the choices of inputs W for any set W. The (2,2,2)-Bell scenario

Table 1. An empirical model for the (2,2,2)-Bell scenario.

	(0, 0)	(0, 1)	(1, 0)	(1, 1)
(a_1, b_1)	1/2	0	0	1/2
(a_1, b_2)	3/8	1/8	1/8	3/8
(a_2, b_1)	3/8	1/8	1/8	3/8
(a_2, b_2)	1/8	3/8	3/8	1/8

depicted in table 1 indeed satisfies this so-called *no-signalling* condition, since, for instance:

$$d_{\{a_1,b_1\}}|_{\{a_1\}}(0) = d_{\{a_1,b_2\}}|_{\{a_1\}}(0) = \frac{1}{2}.$$

Now, a system is said to be *non-contextual* if there exists a joint probability distribution over X which correctly restricts to all of the d_U's. It can be shown that the example of figure 1 actually is *contextual*, i.e. such a global probability distribution cannot be defined. This is related to the formal definition of a sheaf. A *sheaf* is a presheaf $\mathcal{F} : \mathcal{C}^{op} \to \mathbf{Sets}$ s.t. for every pair of objects U, V of $\mathcal{C}$, the restriction morphisms from $\mathcal{F}(U)$ and $\mathcal{F}(V)$ to the intersection $\mathcal{F}(U \cap V)$ agree, i.e. for $h_U \in \mathcal{F}(U)$ and $h_V \in \mathcal{F}(V)$, we have:

$$h_U|_{U\cap V} = h_V|_{U\cap V}. \tag{2.3}$$

We say that h_U and h_V are *compatible* (or consistent) whenever (2.3) is satisfied.

In realistic experiments, the no-signalling condition does not usually hold; this can be due to unsharpness of the instruments [16] or simply finiteness of the measurements [15,16]. As a result, different frameworks have been developed to study contextuality in the presence of signalling. Examples of these are the Contextuality-by-Default framework [15] and the signalling fraction of the sheaf theoretic model [16], both of which create a measure of the signalling property of the system. A signalling system is then said to be contextual if the amount of signalling is not enough to make the system 'classically explainable'. In sheaf-theoretic terminology, if every pair of sections in an empirical model satisfies the compatibility condition of (2.3), then the empirical model is said to be *no-signalling* or *consistent*. If the sections of an empirical model correspond to the statistics of a system, then the no-signalling condition is exactly the compatibility condition of (2.3). Given an empirical model e, which is not necessarily compatible, we define the *no-signalling fraction* $\mathsf{NSF} \in [0, 1]$ as the maximal possible value of λ across all of the decompositions of e:

$$e = \lambda \cdot e_{\mathrm{NS}} + (1 - \lambda) \cdot e' \tag{2.4}$$

where e_{NS} is a no-signalling empirical model (the multiplication here is understood as point-wise multiplication), and e' can be any empirical model. We then define the *signalling fraction* as:

$$\mathsf{SF} = 1 - \mathsf{NSF}. \tag{2.5}$$

This can be seen as the degree of incompatibility of an empirical model, as it measures the departure from a no-signalling, and hence locally compatible model.

In [17–19,32], this formulation of contextuality has been extended to scenarios where structured signalling is allowed, first by allowing sequential operations in [32], then by allowing *definitite causal orders* [17,19] and even *indefinite causal structures* [17,18]. In all of these studies, the introduction of causality was done by relaxing the symmetry property of the compatibility relation on X. In causal scenarios, a being compatible with b will mean that measuring a can (potentially) influence the outcome of the measurement made by b; we will write $a \preceq b$. For example, starting from the same set of measurements as the (2,2,2)-Bell scenario, we can impose

Table 2. An empirical model for a causal scenario with $U = \{a_1, a_2\} \preceq V = \{b_1, b_2\}$.

	$(0, 0)$	$(0, 1)$	$(1, 0)$	$(1, 1)$
(a_1, b_1)	0	6/13	0	7/13
(a_1, b_2)	24/65	6/65	7/13	0
(a_2, b_1)	23/65	0	14/65	28/65
(a_2, b_2)	23/260	69/260	42/65	0

the conditions:

$$a_1 \preceq b_{1,2} \quad \text{and} \quad a_2 \preceq b_{1,2}$$

in order to model the scenario where Alice can do her measurements before Bob. We can subsequently define the probability distributions as before, with the previous no-signalling condition being replaced by the *causality* condition; i.e. if U, V and V' are sets of measurements such that every element $a \in U$, $b \in V$ and $b' \in V'$, we have $a \preceq b$ and $b \preceq b'$ (shorthanded to $U \preceq V, V'$), then:

$$d_{U \cup V}|_U(\underline{o}_U) = d_{U \cup V'}|_U(\underline{o}_U)$$

for any joint outcome $\underline{o}_U$ over U. An example of such a causal scenario is shown in table 2. We can indeed see that for instance:

$$d_{\{a_1,b_1\}}|_{\{a_1\}}(0) = d_{\{a_1,b_2\}}|_{\{a_1\}}(0) = \frac{6}{13}$$

and

$$d_{\{a_2,b_1\}}|_{\{a_2\}}(0) = d_{\{a_2,b_2\}}|_{\{a_2\}}(0) = \frac{23}{65}$$

but:

$$0 = d_{\{a_1,b_1\}}|_{\{b_1\}}(0) \neq d_{\{a_2,b_1\}}|_{\{b_1\}}(0) = \frac{37}{65}$$

and

$$\frac{59}{65} = d_{\{a_1,b_2\}}|_{\{b_2\}}(0) \neq d_{\{a_2,b_2\}}|_{\{b_2\}}(0) = \frac{191}{260}.$$

Therefore, we have $U \preceq V$ but not the converse (i.e. $V \not\preceq U$). Contextuality is, as before, defined as the impossibility of finding a global probability distribution over X which marginalizes to the original probability distributions. Similar to the signalling fraction, given a scenario endowed with an asymmetric compatibility condition, we define its *causal fraction* CausF as the minimal λ across all of the decompositions:

$$e = \lambda \cdot e_{\text{caus}} + (1 - \lambda) \cdot e' \tag{2.6}$$

where e_{caus} is an empirical model consistent with the imposed compatibility relation.

The above notions are defined in the sheaf-theoretic framework of contextuality. In §3, we argue that they also provide a straightforward interpretation in terms of linguistic phenomena.

3. Methodology

We use sheaf theory to model the incremental process of reading and parsing, according to a popular theory known as parallel-ranking [22,33,34]. In this theory, as a human subject reads on, they keep track of and rank all possible local structures. These ranks are used to prune the tree of possibilities, by selecting the structures with the highest 'ranks'. By analogy with sheaf theoretic models of quantum mechanics, we take the measurements to be words, their outcomes to be their grammatical structures and the measurement contexts to be linguistic contexts.

By analogy with the causal models, the compatibility relation is used to encode the linear order of the words in a sentence. The causality condition, if satisfied, then states that the process of building parses is purely incremental; as a new word is read, the reader will simply extend the existing graphs and assign probabilities to the new structures in a consistent way. In this section, we formalize these intuitions.

(a) Sheaf-theoretic models of garden-path sentences

Given a garden-path sentence g, we denote its vocabulary by the set $\Sigma(g)$ and refer to the monoid generated over it as $\Sigma(g)^*$; this monoid will be the base category $\mathcal{C}$ of our presheaf. Elements of $\Sigma(g)^*$ are seen as strings of words, i.e. phrases with elements in $\Sigma(g)$, the monoid multiplication is concatenation, and its unit is the empty string. We endow this monoid with a partial order $- \leq - \subseteq \Sigma(g)^* \times \Sigma(g)^*$, and use it to denote the prefix order between the phrases, defined below:

$$m_1 \leq m_2 \iff m_1 \text{ is the initial subphrase of } m_2.$$

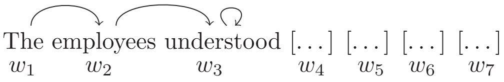
As an example, consider the garden-path sentence $g =$ 'The employees understood the contract would change'. Its vocabulary is the following set:

$$\Sigma(g) = \{\text{employees, understood, the, contract, would, change}\}. \tag{3.1}$$

The monoid $\Sigma(g)^*$ contains all its subphrases and the following prefix orders between them:

$$\text{employees} \leq \text{employees understood} \leq \text{employees understood contract}$$
$$\text{understood} \leq \text{understood the} \leq \text{understood the contract}.$$

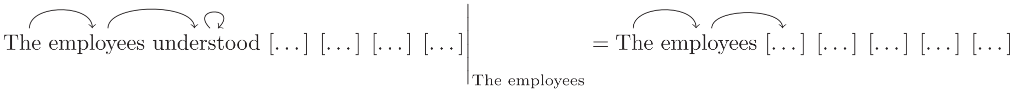
To each $m \in \Sigma(g)^*$, we assign any *linguistic structure* that it might carry. This can be syntactic structure, semantic structure (e.g. its degree of plausibility), or pragmatic structure (e.g. the intentions or questions behind it). In this paper, our focus is on syntactic structure, and we choose to work with dependency relations [35]. These relations encode syntactic dependencies between words in a phrase or sentence. For instance, when a noun such as 'employee' is modified by an article, such as 'the', then the article depends on the noun. The modified word will then be called the *head* of the modifying word; this head is unique for any given word in a phrase. The set of all possible heads will be our set of outcomes. In a sentence of length n, each word can have n different possible heads. Hence, our set of outcomes is $O = \{1, \ldots, n\}$ where n is the length of the sentence g.[2] The maps $s : m \to O$ of the event presheaf assign syntactic structures to subphrases of a sentence. As an example, consider the sentence fragment 'The employees understood'; a possible dependency parse for this fragment is as follows, where an arrow $w_i \to w_k$ indicates w_k is the head of w_i:



The function s: 'The employees understood' $\to O = \{1, \ldots, 7\}$ assigns this structure to the sentence fragment as follows:

$$s = \{\text{The}_1 \mapsto 2; \text{employees}_2 \mapsto 3; \text{understood}_3 \mapsto 3\}.$$

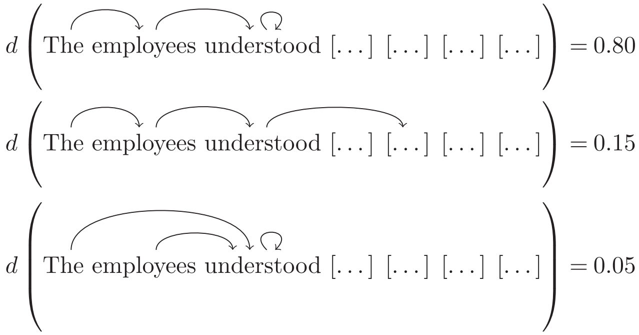
The restriction maps of our presheaf $\mathcal{E}$ are defined as in formula (2.1). For example, the restriction of the above grammatical structure from 'The employees understood' to 'The employees' is:


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

The function corresponding to it is $s|_{\text{The employees}} = \{\text{The}_1 \mapsto 2; \text{employees}_2 \mapsto 3\}$. The action of the distribution monad in turn sends sentence fragments to the set of *all* of the distributions over all

[2] If we were to consider sentences of any length, we could also take $O = \mathbb{N}$.

of their grammatical structures. For example, for $m_1 =$ 'The employees' and $m_2 =$ 'The employees understood', we could have $d \in \mathcal{D}_{\mathbb{R}_+}\mathcal{E}(m_2)$ s.t.:


d (The employees understood [...] [...] [...] [...]) = 0.80
d (The employees understood [...] [...] [...] [...]) = 0.15
d (The employees understood [...] [...] [...] [...]) = 0.05

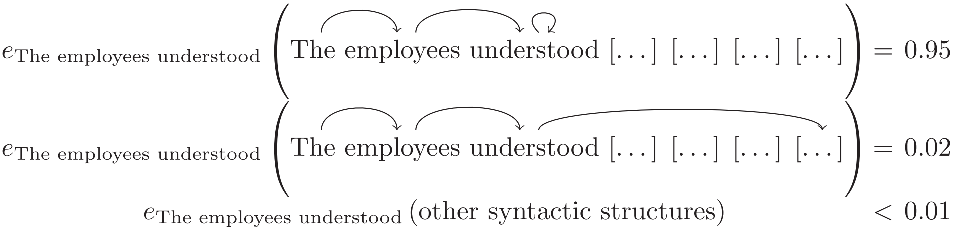
Their restriction morphisms are defined as follows. For $m_1 = w_1 \ldots w_n$ and $m_2 = w_1 \ldots w_n w_{n+1} \ldots w_k$ (i.e. $m_1 \leq m_2$ in $\mathcal{C}$), the restriction morphism from $\mathcal{D}_{\mathbb{R}_+}\mathcal{E}(m_2) \to \mathcal{D}_{\mathbb{R}_+}\mathcal{E}(m_1)$ takes any $d \in \mathcal{D}_{\mathbb{R}_+}\mathcal{E}(m_2)$ to the probability distribution s.t.:

$$d|_{m_1}(o_1 \ldots o_n) = \sum_{(o_{n+1} \ldots o_k) \in O^{k-n}} d(o_1 \ldots o_n o_{n+1} \ldots o_k). \tag{3.2}$$

In order to mimic the human reading process, we consider a *sequence* of empirical models, based on collections of subphrases $\{\mathcal{M}_i\}_{1 \leq i \leq n-1}$, where n is the length of the sentence. Elements of $\mathcal{M} = \{m_i, m_{i+1}\}$ are strings of lengths i and $i+1$, respectively, such that $m_i \leq m_{i+1} \leq g$. The sequences represent the evolution of the linguistic contexts. For instance, for $g =$ 'The employees understood the contract would change', we have the following sequence:

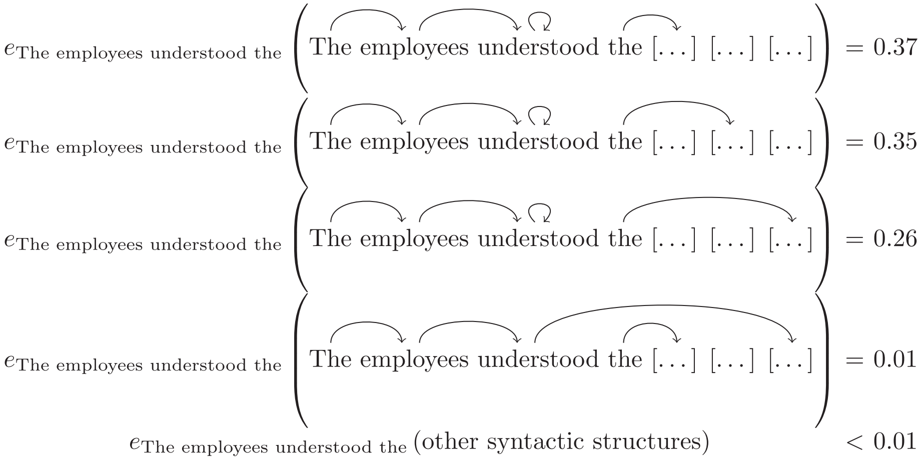
$$\begin{aligned}
\mathcal{M}_1 &= \{\text{The, The employees}\}\\
\mathcal{M}_2 &= \{\text{The employees, The employees understood}\}\\
\mathcal{M}_3 &= \{\text{The employees understood, The employees understood the}\}\\
&\ \ \vdots\\
\mathcal{M}_6 &= \{\text{The employees understood the contract would,}\\
&\qquad \text{The employees understood the contract would change}\}
\end{aligned}$$

Recall that each of the empirical models consists of a pair of sections, i.e. a pair of probability distributions over grammatical structures; as we will see in §3c, these are obtained empirically. For example, a realization of $\mathcal{M}_3 = \{e_{\text{The employees understood}}, e_{\text{The employees understood the}}\}$ in the above sequence is:


e_The employees understood (The employees understood [...] [...] [...] [...]) = 0.95
e_The employees understood (The employees understood [...] [...] [...] [...]) = 0.02
e_The employees understood (other syntactic structures) < 0.01

(3.3)

and:



$$
\begin{aligned}
&e_{\text{The employees understood the}}\left(\text{The employees understood the } [\ldots]\ [\ldots]\ [\ldots]\right) = 0.37\\
&e_{\text{The employees understood the}}\left(\text{The employees understood the } [\ldots]\ [\ldots]\ [\ldots]\right) = 0.35\\
&e_{\text{The employees understood the}}\left(\text{The employees understood the } [\ldots]\ [\ldots]\ [\ldots]\right) = 0.26\\
&e_{\text{The employees understood the}}\left(\text{The employees understood the } [\ldots]\ [\ldots]\ [\ldots]\right) = 0.01\\
&e_{\text{The employees understood the}}(\text{other syntactic structures}) < 0.01
\end{aligned}
\tag{3.4}
$$

In this setting, the compatibility condition says that the maps that assign degrees of likelihood to syntactic structures of a common prefix of a subphrase, are compatible—in other words—agree with each other. This does not usually happen in data collected statistically. In the case of garden-path sentences, the amount of incompatibility is exacerbated, as the probabilities associated with certain syntactic structures dramatically change when entering the critical region. In other words, we should expect high signalling/non-causal fractions at the boundary of the critical region.

(b) Analysis of the linguistic empirical models

In this section, we explore the properties of the linguistic models introduced above. We start by observing that the linguistic models can be seen as both a contextuality and a causality scenario. To see this, we first note that for each empirical model, one context includes exactly one less word than the other. As a result, we can w.l.o.g see empirical models as an $\{m, mw\}$ scenario. For example, in the empirical model $\mathcal{M}_3$ of §3a, we have m = 'The employees understood' and w = 'the'. As a result, the compatibility relation of the linguistic models can be expressed as follows:

1. A symmetric relation analogous to measurements that can be simultaneously measured. In this case, we interpret m and w as being compatible, which means that m and w are somewhat parsed independently. In this interpretation, we have a situation similar to contextuality scenarios;
2. An asymmetric relation analogous to causal scenarios, where the compatibility relation read as $m \preceq w$.

Both of these interpretations are possible, and very much related. However, the meanings of the quantity SF are subtly different. In the symmetric interpretation, SF quantifies how consistent the two probability distributions are. On the other hand, in the causal interpretation, the causal fraction is also a measure of the departure from a causal model imposed by the linear *reading* order; in other words, a high signalling fraction is evidence for the fact that the process of assigning a syntactic structure to a particular subphrase is not incremental, but instead should require information coming from the words situated *after* the phrase under consideration. For the rest of this paper, we will mostly focus on the signalling view of the models.

In addition, not all of the measurement scenarios are capable of hosting contextuality, and the structure of the $\mathcal{M}$ can be used to identify this. One can associate a simplicial complex with every $\mathcal{M}$ straightforwardly by defining for each $C \in \mathcal{M}$ with $|C| = n$ an $n+1$-simplex, with a vertex for each local measurement, and identifying the faces of $C, C' \in \mathcal{M}$ corresponding to $C \cap C'$. Vorob'ev's theorem [36] then implies that if an empirical model is defined over a measurement context for which the topology of $\mathcal{M}$ does not contain a cycle, then it is non-contextual [37]. Since the linguistic contexts in each of our measurement scenarios are totally ordered, the geometric representations of the $\mathcal{M}_i$'s do not contain any cycle and therefore cannot, by design, be contextual.

Finally, computing the signalling/causal fractions in a generic empirical model is not a trivial task, as it requires finding a solution to a linear optimization problem [16]. However, given the specific structure of our empirical models, it is possible to find an expression of the signalling fraction SF that can be calculated efficiently. This expression is shown in (A 10) and its proof can be found in appendix A.

Proposition 3.1. *The signalling fraction can be computed via the following equation:*

$$\mathsf{SF} = 1 - \sum_{o} \min(e_{mw}|_m(o), e_m(o)). \tag{3.5}$$

We argue that the signalling fraction can be seen as a measure of difficulty when assigning grammatical structure, in other words, parsing difficulty. This is motivated by the fact that SF can be seen as a measure of distance between probability distributions observed at different stages of the sentence. Therefore, the higher the syntactic fraction, the more a reader will have to readjust their mental representation of the grammatical structure. We can even say that since the contexts $m_i, m_{i+1} \in \mathcal{M}_i$ only differ by a single word, the signalling fraction of the empirical model e_i becomes related to the difficulty of understanding the extra word. For example, for the empirical model $\mathcal{M}_3$ defined in (3.3) and (3.4), we obtain a signalling fraction of $\mathsf{SF}_3 = 0.05$, hence showing that the word 'the' at the end of the fragment 'The employees understood the' is not difficult to parse. On the other hand, if we calculate the signalling fraction for the empirical model $\mathcal{M}_5$ (see figure 2 in appendix B), for the signalling fraction we have $\mathsf{SF}_5 = 0.79$, which reflects the fact the parsing the word 'would' is quite difficult.

(c) Data collection using deep neural network language models

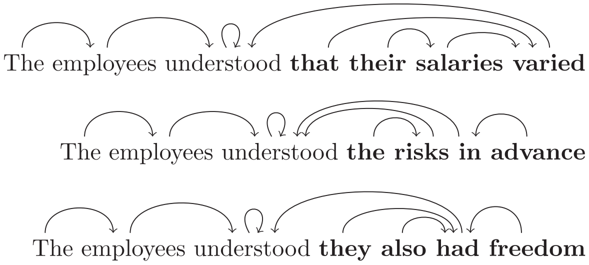
In order to obtain probabilities of different syntactic structures of subphrases of a garden-path sentence, we need to input these subphrases to an automatic dependency parser and we use the state-of-the-art dependency parser `spaCy` [38]. As `spaCy` only assigns syntactic structures to completed sentences, we need to first come up with completions of each subphrase under consideration. To this end, we use the *masking* tool of another state-of-the-art NLP tool: the Bidirectional Encoder Representations from Transformers language model, also known as `BERT` [39,40]. `BERT` is trained on a word-in-context prediction task, that is, given a sequence of words where one or more words are *masked*, the neural network is trained to predict the values of those masks and provide a degree of likelihood for each of its predictions. We use `BERT` to obtain completions of phrases and their probabilities according to the following steps:

1. Given a subphrase of a garden-path sentence, we turn it into a complete sentence by masking all of the remaining words of the g, see figure 1 for an example.
2. `BERT` provides a list of predictions of the completion of the subphrases[3] and a *logit* score s for each of these predictions, which is meant to rate the likelihood of each prediction. The common practice in NLP is to use the logistic function $p = e^s/(1 + e^s)$ to turn these scores into probabilities.
3. We then use `spaCy` to parse each of the predictions provided by `BERT`. In order to obtain the grammatical structure of the specific subphrase we are working with, we restrict the

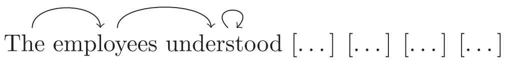
[3] These predictions are not always words and may include punctuations. To avoid complications, we only include the text predictions by dropping the punctuation and renormalizing the probability distribution.

full parse to the words included in that subphrase using the restriction maps from the presheaf $\mathcal{E}$ (see equation (2.1)).

4. The probability of each such (partial) parse is obtained by summing up all the BERT-prediction probabilities which restrict to the same parse. For example, consider the predictions for the continuations of 'The employees understood':


The employees understood **that their salaries varied**
The employees understood **the risks in advance**
The employees understood **they also had freedom**

These lead to the same partial parse when restricted to the context 'The employees understood', namely:



The resulting probability distributions are subsequently used in empirical models to calculate signalling fractions.

We work with different variations of BERT and spaCy. For BERT, first and foremost we worked with its original version bert-base-cased, which has 110 million parameters and was trained on Toronto Book Corpus and the English Wikipedia, both of which are cased, i.e. distinguish between lower and upper case letters. Uncased versions of the same algorithm exist and were developed for purposes of cross-lingual learning. Secondly, we worked with bert-large-cased, which has 340 million parameters. Finally, we worked with distilBERT, which only has 40% of the parameters of the original BERT model, but runs 60% faster while preserving 95% of its performance accuracies in language understanding tasks. For spaCy, we worked with its small and large versions, called en_core_web_sm and en_core_web_lg. Both models are trained using convolutional neural networks on web text consisting of blogs, news and comments. The difference between them is that the small version does not have any word vectors whereas the large one has a word vector table with 500k unique 300-dimensional vectors. We also worked with a newer version of spaCy,en_core_web_trf, which has no word vectors but is trained using state-of-the-art transformer-based neural networks. We compare the values of SF and the accuracies of their predictions between the different versions of these tools. The consensus in the machine learning community is that models trained on a larger corpus with a larger set of parameters, including the dimensionality of vectors, should be more accurate. We work with their different variants to obtain a better understanding of the robustness of our results.

4. Results

We work with the datasets of Sturt *et al.* [21] and the dataset of Grodner *et al.* [22]. Each garden-path sentence of each of these datasets is paired with its disambiguated version. The Sturt *et al.* dataset consists of 32 ambiguous NP/S, and 32 ambiguous NP/Z sentences. Taking the disambiguated versions of each sentence into account, this gives a total of 128 sentences. Sturt *et al.* divide each of their sentences, whether garden-path or unambiguous, into four regions, see table 3*a*, where the penultimate region is the *critical region*. They report reading times for each of the regions; these reading times are then averaged over the different sentences of the same type (i.e. NP/S ambiguous, NP/S unambiguous, NP/Z ambiguous or NP/Z unambiguous). The dataset of Grodner *et al.* consists of 40 ambiguous NP/S and 40 ambiguous NP/Z sentences and their disambiguated variants, i.e. a total of 160 sentences. Similar to the Sturt *et al.* dataset, each sentence is divided into regions, but in this case, the regions are different for the different types of sentences, see table 3*b*. Grodner *et al.* reports an average word-by-word reading times for each region, averaged again across the different sentences of the same type.[4] The reading times of these dataset are in table 10 appendix C(a).

The data collection procedures differed in [21,22] and one can argue that the reading times they studied are fundamentally different. We did not use standard methods to unify these datasets to avoid the systematic error caused by the human/computer reaction times. These are induced by the change of stimuli (i.e. when a new word or phrase is displayed on screen). Assuming that this time is constant for a given participant, it occurs once per region in the region-by-region setting of Sturt *et al*., but multiple times per region in the word-by-word setting of Grodner *et al.* With regard to this, what we are investigating is the assumption that only the reading difficulty of a given word is affected by SF, all other factors are kept constant. Then, in the case of the word-by-word setting, we are mimicking the averaging procedure of Grodner *et al*. [22] and study the correlation between the average of the word-by-word SF, across a given region, and all the sentences of this type. On the other hand, in the case of the region-by-region setting, we are investigating the correlation between the *sum* of the word-by-word SF in a given region and the corresponding region's reading time.

The empirical models obtained for all of the variants of the BERT and spaCy tools for all of the sentences of these datasets are available online at [41].

(a) Correlations between SF and reading times

We test the degree of correlation between SF and reading times and find a strong positive correlation. The resulting Pearson's ρ coefficients and associated p-values are in table 4. For both datasets and regardless of the choice of BERT and spaCy, Pearson's coefficients are statistically significantly above 0. They are also fairly high; in particular, for the dataset of Sturt *et al.* [21]. By considering all of the data obtained from all of the BERT and spaCy combinations, we obtain the following linear regression equations:

$$\mathrm{RT}_{\mathrm{Sturt}}(\mathrm{region}) = 295 \sum_{w\in\mathrm{region}} \mathsf{SF}(w) + 664\,\mathrm{ms} \tag{4.1}$$

and

$$\mathrm{RT}_{\mathrm{Grodner}}(w) = 77\,\mathsf{SF}(w) + 381\,\mathrm{ms}. \tag{4.2}$$

The individual linear regression equations for each of the specific BERT/spaCy models combinations were also solved and are provided in appendix C(b) table 11. All coefficients (the individual ones, as well as the ones of (4.1) and (4.2)) were within each other's standard error. We used the individual linear regression models of table 11 for our prediction task, presented in the next paragraph.

(i) Detecting a garden-path effect

Given that SF is correlated with reading times, we expect to observe a difference between SF of the garden-path sentences versus SF of their unambiguous variants, in their critical regions. This is indeed the case, see table 5. As we can see, the average difference in SF is positive. In order to test whether the departures from 0 are significant, in table 6, we do a t-test for the null hypothesis that the differences between SF in garden-path sentences and SF in the corresponding unambiguous

[4] Half of the NP/S and NP/Z sentences (20 sentences each) were labelled as 'modified', as they contained a descriptive noun-phrase attached to the subject of the main verb, whereas the rest of the sentences were labelled as 'unmodified'. This distinction was made since the authors were investigating whether adding modifying noun-phrase would alter the misanalysis difficulty. Since we are not interested in this distinction in the current work, we will treat both (un)modified versions as the same.

Table 3. Example of different regions of different datasets. The critical regions are in italics.

	regions					
(*a*) Sturt *et al.* dataset						
NP/S (ambiguous)	the faithful employees	understood the technical contract	*would be changed*	very soon		
NP/S (unambiguous)	the faithful employees	understood that the technical contract	*would be changed*	very soon		
NP/Z (ambiguous)	because the employees	negotiated the technical contract	*would be changed*	very soon		
NP/Z (unambiguous)	because the employees	negotiated, the technical contract	*would be changed*	very soon		
(*b*) Grodner *et al.* dataset						
NP/S (unmod., ambiguous)	the employees	understood	the contract	*would be changed*		
NP/S (unmod., unambiguous)	The employees	understood	that	the contract	*would be changed*	
NP/S (mod., ambiguous)	the employees	who initiated the strike	understood	the contract	*would be changed*	
NP/S (mod., unambiguous)	the employees	who initiated the strike	understood	that	the contract	*would be changed*
NP/Z (unmod., ambiguous)	even though the band	left	the party	*went on for* […]		
NP/Z (unmod., unambiguous)	Even though the band	left,	the party	*went on for* […]		
NP/Z (mod., ambiguous)	even though the band	which played funk music	left	the party	*went on for* […]	
NP/Z (mod., unambiguous)	even though the band	which played funk music	left,	the party	*went on for* […]	

Table 4. Pearson's ρ coefficients and associated *p*-values (in brackets) between SF and reading times. The statistically significant *p*-values are in italics. Only the *p*-values marked with † are not statistically significant after Bonferroni correction.

		BERT model		
		`distilbert`	`bert-base`	`bert-large`
(*a*) Sturt *et al.* dataset				
`spaCy` model	`en_core_web_sm`	0.64 (*0.008*)	0.80 (*0.0002*)	0.79 (*0.0003*)
	`en_core_web_lg`	0.63 (*0.009*)	0.79 (*0.0003*)	0.78 (*0.0003*)
	`en_core_web_trf`	0.67 (*0.004*)	0.78 (*0.0004*)	0.76 (*0.0006*)
(*b*) Grodner *et al.* dataset				
`spaCy` model	`en_core_web_sm`	0.55 (*0.0004*)	0.35 (0.03†)	0.56 (*0.0002*)
	`en_core_web_lg`	0.52 (*0.0008*)	0.37 (0.02†)	0.48 (*0.002*)
	`en_core_web_trf`	0.53 (*0.0006*)	0.36 (0.02†)	0.44 (*0.006*)

Table 5. Average difference of SF between garden-path sentences and their unambiguous variants with standard deviation. Note: for the dataset of Sturt *et al*. [21], the sum of SF over the critical regions are quoted, while for the dataset of Grodner *et al*. [22], the mean of SF is shown.

		BERT model		
		`distilbert`	`bert-base`	`bert-large`
(*a*) Sturt *et al.* dataset				
`spaCy` model	`en_core_web_sm`	0.07 ± 0.18	0.15 ± 0.16	0.16 ± 0.17
	`en_core_web_lg`	0.03 ± 0.14	0.15 ± 0.18	0.17 ± 0.17
	`en_core_web_trf`	0.18 ± 0.20	0.29 ± 0.18	0.29 ± 0.19
(*b*) Grodner *et al.* dataset				
`spaCy` model	`en_core_web_sm`	0.02 ± 0.11	0.03 ± 0.10	0.05 ± 0.11
	`en_core_web_lg`	0.007 ± 0.08	0.05 ± 0.08	0.06 ± 0.10
	`en_core_web_trf`	0.06 ± 0.06	0.07 ± 0.06	0.06 ± 0.07

variants are on average 0. As we can see, the *p*-values are quite low in most cases, and they approach 0 for the `BERT` and `spaCy` models with a larger set of parameters.[5]

(ii) Garden-path effect predictions

We used each of the individual regression equations of table 11 to predict the reading times resulting from SF. We predict the garden-path effect by taking the difference between predictions for garden-path sentences and their unambiguous variants over the critical region. For all `BERT` and `spaCy` models, our SF predictions underestimate the garden-path effects, in particular for NP/Z sentences. However, the predicted effects increase as the `BERT` and `spaCy` models get larger. The average predictions are depicted in table 7 (see also figure 3 in appendix C(c) for more details). On the whole, `bert-base` and `bert-large` provided better predictions than `distilbert`; similarly the `en_core_web_trf` outperformed other variants of `spaCy`. Some of the combinations using smaller models did not provide accurate predictions and in fact

[5] One may also note that, although the results appear to be better as the `BERT` and `spaCy` models get better, the highest correlation and, respectively, the lowest *p*-value are not necessarily obtained for the combination using both of the the `BERT` and `spaCy` variants with the largest number of parameters. This could be due to statistical variations. This remark will also apply to the subsequent results as well.

Table 6. *p*-values associated with the *t*-tests evaluating whether the unambiguous and ambiguous SF are the same. The lower these numbers, the more significant the results. The statistically significant results are highlighted in italics.

		BERT model		
		`distilbert`	`bert-base`	`bert-large`
(*a*) Sturt *et al.* dataset				
`spaCy` model	`en_core_web_sm`	*0.005*	$<10^{-8}$	$<10^{-9}$
	`en_core_web_lg`	0.07	$<10^{-7}$	$<10^{-10}$
	`en_core_web_trf`	$<10^{-9}$	$<10^{-18}$	$<10^{-17}$
(*b*) Grodner *et al.* dataset				
`spaCy` model	`en_core_web_sm`	0.13	*0.01*	*0.0004*
	`en_core_web_lg`	0.46	$<10^{-5}$	$<10^{-5}$
	`en_core_web_trf`	$<10^{-10}$	$<10^{-13}$	$<10^{-9}$

Table 7. Average predicted garden-path effects for NP/S and NP/Z sentences in milliseconds (the type of sentences is shorthanded in brackets: S for NP/S and Z for NP/Z).

		BERT model		
		`distilbert`	`bert-base`	`bert-large`
(*a*) Sturt *et al*. The human garden-path effect is 87 ms for NP/S sentences and 400 ms for NP/Z sentences				
`spaCy` model	`en_core_web_sm`	4.41(S);30.0(Z)	33.0(S);69.6(Z)	41.6(S); 65.4(Z)
	`en_core_web_lg`	−1.89(S);17.5(Z)	33.5(S);65.8(Z)	47.1(S); 63.4(Z)
	`en_core_web_trf`	35.2(S);44.9(Z)	60.9(S);110(Z)	62.6(S); 93.6(Z)
(*b*) Grodner *et al*. The human garden-path effect is 21 ms for NP/S sentences and 53.5 ms for NP/Z sentences				
`spaCy` model	`en_core_web_sm`	−0.51(S);3.90(Z)	−0.30(S);4.41(Z)	0.73(S); 9.36(Z)
	`en_core_web_lg`	−1.00(S);2.07(Z)	2.25(S);4.39(Z)	2.15(S); 8.52(Z)
	`en_core_web_trf`	2.73(S);4.71(Z)	2.62(S);5.57(Z)	1.90(S); 7.41(Z)

predicted the wrong trend (i.e. predicted that garden-path sentences were read *faster* than their unambiguous analogues).

(iii) Comparison of NP/S and NP/Z predictions

Now that we have predicted the garden-path effect for NP/S and NP/Z sentences, we are interested to see whether SF distinguishes between them. We performed *t*-tests to verify this hypothesis. The *p*-values are in table 8. For most of the `BERT` and `spaCy` variants, there is a statistically significant difference in reading time predictions[6] (apart from a couple of exceptions in both datasets). This shows that SF is able to witness the increase of difficulty induced by having an NP/Z ambiguity type. As before, `bert-base` or `bert-largse` and the `en_core_web_trf` variants distinguished the NP/S and NP/Z sentences more accurately.

[6] In the case of the Sturt *et al.* dataset, most of the *p*-values are no longer significant after Bonferroni correction with significance level $\alpha = 0.05$; however, the *p*-values do remain low and only one entry of the table will no longer be statistically significant after Bonferroni correction at significance level $\alpha = 0.1$, which may be more relevant in this study anyway given the noisiness of the data.

Table 8. *p*-values associated with the *t*-test evaluating whether the garden-path effects obtained from SF for NP/S and NP/Z are sampled from the same distribution. The statistically significant *p*-values are highlighted in italics. Only the *p*-values marked with † are not statistically significant after Bonferroni correction.

		BERT model		
		`distilbert`	`bert-base`	`bert-large`
(*a*) Sturt *et al.* dataset				
`spaCy` model	`en_core_web_sm`	**0.03**†	*0.01*	0.09
	`en_core_web_lg`	**0.02**†	**0.04**†	0.24
	`en_core_web_trf`	0.39	*0.0001*	*0.01*
(*b*) Grodner *et al*. dataset				
`spaCy` model	`en_core_web_sm`	**0.04**†	*0.002*	*0.001*
	`en_core_web_lg`	**0.02**†	0.09	*0.001*
	`en_core_web_trf`	**0.03**†	*0.0006*	$<\mathbf{10}^{-5}$

Table 9. Optimal (averaged) garden-path effect predictions. The best predictions are denoted in italics.

		prediction (ms)		
		SF	SP [13]	observed (ms)
Sturt *et al.*	NP/S	*62.6*	24	87
	NP/Z	*110*	30	400
Grodner *et al.*	NP/S	2.73	*7*	21
	NP/Z	8.52	*10*	53.5

(b) Comparison with surprisal

Schijndel & Linzen [13] generated a linear regression model of reading times as a function of surprisal, and used it to predict garden-path effects. Similar to us, they observed that surprisal underestimates the garden-path effects, but different from us, they could not distinguish between the NP/S and NP/Z sentences [13,26,27].

Although statistical tests are not quoted in [13], our model clearly outperforms the results obtained from surprisal and our predictions for NP/Z sentences are significantly higher than for NP/S sentences, see table 9. The SF values are closer to the observed human garden-path effects in the Sturt *et al.* dataset [21], and in the Grodner *et al.* dataset [22], they are comparable to the predictions provided by surprisal. For instance, the SF values of the NP/S sentences of [21] predict a slowdown of about 63 ms, whereas surprisal only predicts a slowdown of 24 ms at best (the human effect is 87 ms).

Even though our usage of SF stemmed from similar motivations to those for surprisal, it is not clear whether they are mathematically related. The reason for the better performance of SF is that surprisal, as used in [13], mostly focuses on lexical items, whereas for the SF quantity described here, syntactic structures are first-class citizens. Only very recently (this year), the role of syntactic structure in conjunction with surprisal has come to light: in [28], it was shown that syntactic surprisal performs slightly better than pure lexical surprisal, but still falls short when distinguishing NP/S from NP/Z and the differences in garden-path effects. The results of [28] and the ones presented here motivate the hypothesis that syntactic structures are the main deciding factor in the difficulty of garden-path sentences. Another aspect of our work, which may have led to more accurate results, is that our model is able to take long-distance dependencies into account, whereas surprisal is not.

5. Conclusion and discussion

In this work, we constructed a sheaf-theoretic model of human sentence processing, inspired by similar models in quantum contextuality and causality. The base category of our sheaf was a partial order category of sentence fragments, where the signalling and causal fractions coincided. We applied our model to formalize and reason about a challenge known as the garden-path effect, which happens in certain sentences with local syntactic ambiguity. Garden-path sentences have higher reading times in comparison to their locally unambiguous versions. We showed that the signalling/causal fraction correlated well with human reading times. Using this correlation, we predicted reading times with a par (in the Grodner *et al.* [22] dataset) and better (in the Sturt *et al.* [21] dataset) accuracies than recent research in NLP that uses an information-theoretic measure called 'surprisal'. Furthermore, we could observe significantly different predictions for easy and hard garden-path sentences; to date, surprisal-based models have not been able to do so. The better performance of the sheaf-theoretic model is that it takes both linguistic structure and lexical statistics into account, whereas surprisal-based models are only able to focus on one or the other.

Factors other than syntactic structure and lexical statistics are argued for when analysing garden-path sentences, e.g. semantic plausibility [42,43] and pragmatic concerns [21]. In this work, we only considered syntax. However, sheaf theory offers tools for modelling semantic plausibility, e.g. via composition with a Boolean truth-value functor. At this time, it is not clear to us how pragmatics concerns can be modelled. Furthermore, we only modelled forward-looking human processes. Psycholinguistic research shows that humans require backtracking in order to process garden-path sentences. We believe backtracking can be formalized in our sheaf-theoretic model, e.g. by using different compatibility relations. A third limitation of our work has been the differences between the datasets available to us. Indeed, the reading times used for doing the linear regression were only averages of reading times across different sentences and different participants; and this caused discrepancies in our results. We plan to use more detailed datasets such as a recent one released in [27] in order to confirm our results.

Data accessibility. This article has no additional data.

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. D.W.: conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, visualization, writing—original draft; M.S.: conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, writing—original draft, writing—review and editing.

Both authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. DW's research is supported by the Engineering and Physical Sciences Research Council (grant no. EP/S021582/1). MS's research is supported by a Royal Academy of Engineering Senior Fellowship/Research Chair (grant no. RCSRF2122-14-152).

Acknowledgements. We would like to thank Samson Abramsky, Ehtibar Dzhafarov, Kin Ian Lo, Shane Mansfield, Wing Yee Chow, Yasutada Sudo, and Richard Breheny for discussions.

Appendix A. Proof of proposition 3.1

Each of our contexts includes exactly one less word than the next one. As a result, given a pair of successive contexts, we can w.l.o.g consider the 2-context scenario as an $\{m, mw\}$ scenario. The no-signalling condition of the model is then as follows:

$$e_{mw}|_m = e_m. \tag{A 1}$$

Given an arbitrary 2-context empirical model as above, we want to find the following decomposition:

$$e = \mathsf{NSF} \cdot e_{\mathrm{NS}} + \mathsf{SF} \cdot e' \tag{A 2}$$

where e_{NS} is the maximum possible across all such decompositions. Let us assume, as above, that m is a single 'observable' with possible outcomes in $O^m = \{0, \ldots, n^{|m|} - 1\}$, where $|m|$ is the number of words in the context m. Our first goal is to find a distribution for m in e_{NS}, which satisfies the following for all $o_i \in O^m$:

$$\mathsf{NSF}e_{\mathrm{NS},m}(o_i) \leq \min(e_{mw}|_m(o_i), e_m(o_i)). \tag{A 3}$$

From the above, it follows that

$$\sum_{o_i} \mathsf{NSF}e_{\mathrm{NS},m}(o_i) = \mathsf{NSF} \leq \sum_{o_i} \min(e_{mw}|_m(o_i), e_m(o_i)). \tag{A 4}$$

One can always construct an empirical model e_{NS} s.t.:

$$\sum_{o_i} \min(e_{mw}|_m(o_i), e_m(o_i))e_{\mathrm{NS}} \leq e. \tag{A 5}$$

In order to see this, first observe that the probability distribution of $e_{\mathrm{NS},m}$ can be constructed by first relabeling the outcomes to o_{i_k} for $0 \leq k \leq n^{|m|} - 1$ s.t. for $N = n^{|m|} - 1$, the following holds:

$$\begin{aligned} \min(e_{mw}|_m(o_{i_0}), e_m(o_{i_0})) &\leq \min(e_{mw}|_m(o_{i_1}), e_m(o_{i_1})) \\ &\leq \cdots \leq \min(e_{mw}|_m(o_{i_N}), e_m(o_{i_N})). \end{aligned} \tag{A 6}$$

Then, we take σ to be $\sum_{o_i} \min(e_{mw}|_m(o_i), e_m(o_i))$ and set:

$$e_{\mathrm{NS},m}(o_{i_0}) = \frac{\min(e_{mw}|_m(o_{i_0}), e_m(o_{i_0}))}{\sigma}. \tag{A 7}$$

We can then inductively define:

$$e_{\mathrm{NS},m}(o_{i_k}) = \min\left(\frac{\min(e_{mw}|_m(o_{i_0}), e_m(o_{i_0}))}{\sigma}, 1 - \sum_{j=0}^{k-1} e_{\mathrm{NS},m}(o_{i_j})\right). \tag{A 8}$$

From this definition, we know that for all k, we have the following:

$$\sigma e_{\mathrm{NS},m}(o_{i_k}) \leq e_{mw}|_m(o_{i_k}), e_m(o_{i_k}). \tag{A 9}$$

In addition, the above forms a valid probability distribution as, if there exists a k s.t. $e_{\mathrm{NS},m}(o_{i_k}) = 1 - \sum_{j=0}^{k-1} e_{\mathrm{NS},m}(o_{i_j})$, then $e_{\mathrm{NS},m}(o_{i'_k}) = 0$ for all $k' > k$ and therefore:

$$\sum_{k} e_{\mathrm{NS},m}(o_{i_k}) = 1. \tag{A 10}$$

Similarly, if for all k, $e_{\mathrm{NS},m}(o_{i_k}) = \min(e_{mw}|_m(o_{i_0}), e_m(o_{i_0}))/\sigma$, then by definition of σ, we also have (A 10). We extend this probability distribution to an empirical model over $\{m, mw\}$, by defining:

$$e_{\mathrm{NS},mw}(o_m, o_w) = e_{mw}(o_m, o_w)\frac{e_{\mathrm{NS},m}(o_m)}{e_{mw}|_m(o_m)}. \tag{A 11}$$

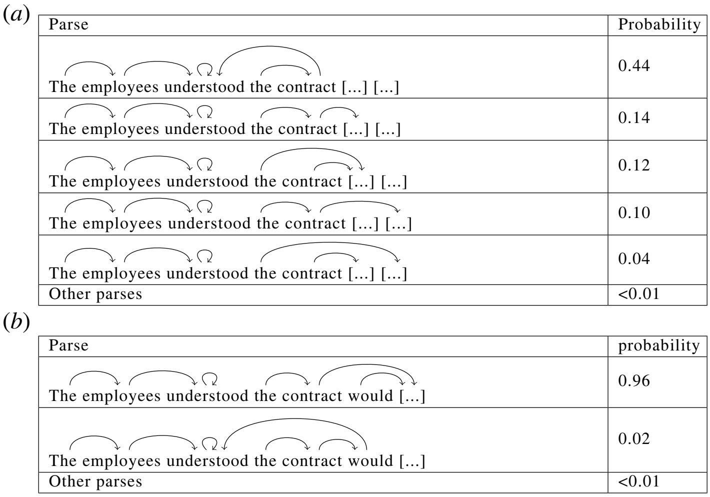
It is now easy to show that $e_{\mathrm{NS},mw}|_m = e_{\mathrm{NS},m}$, and in addition, we have:

$$\sigma e_{\mathrm{NS},mw}(o_m, o_w) = e_{mw}(o_m, o_w)\frac{\sigma e_{\mathrm{NS},m}(o_m)}{e_{mw}|_m(o_m)} \leq e_{mw}(o_m, o_w). \tag{A 12}$$

As a result of the above calculations, the signalling fraction can be computed as follows:

$$\mathsf{SF} = 1 - \sum_{o} \min(e_{mw}|_m(o), e_m(o)). \tag{A 13}$$

Appendix B. Example of an empirical model



(*a*)

Parse	Probability
The employees understood the contract [...] [...]	0.44
The employees understood the contract [...] [...]	0.14
The employees understood the contract [...] [...]	0.12
The employees understood the contract [...] [...]	0.10
The employees understood the contract [...] [...]	0.04
Other parses	<0.01

(*b*)

Parse	probability
The employees understood the contract would [...]	0.96
The employees understood the contract would [...]	0.02
Other parses	<0.01

Figure 2. Example of an empirical model corresponding to $\mathcal{M}_5$ in the sentence 'The employees understood the contract would change' (adapted from the empirical model obtained for 'The faithful employees understood the technical contract would be changed', which can be found in [41]). (*a*) Probability distribution for the context 'The employees understood the contract'. (*b*) Probability distribution for the context 'The employees understood the contract would'.

Appendix C. Details of the garden-path effects

(a) Reading times collected in [21,22]

Table 10. Reading times (in milliseconds) reported in [21,22].

	regions					
(*a*) Sturt *et al*. dataset						
NP/S (ambiguous)	990	1183	877	771		
NP/S (unambiguous)	981	1282	790	768		
NP/Z (ambiguous)	914	1269	1335	848		
NP/Z (unambiguous)	998	1384	935	832		
(*b*) Grodner *et al.* dataset						
NP/S (unmod., ambiguous)	397	467	412	424		
NP/S (unmod., unambiguous)	393	460	431	396	410	
NP/S (mod., ambiguous)	392	415	449	401	419	
NP/S (mod., unambiguous)	398	413	471	393	388	391
NP/Z (unmod., ambiguous)	452	402	382	452		
NP/Z (unmod., unambiguous)	400	452	402	383		
NP/Z (mod., ambiguous)	433	407	464	415	432	
NP/Z (mod., unambiguous)	405	400	494	448	395	

(b) Individual linear regression models

Table 11. Individual linear regression models (in milliseconds).

	distilbert	bert-base	bert-large
(*a*) Sturt *et al.* dataset			
en_core_web_sm	$263(\pm 85)\sum \mathrm{SF} + 690(\pm 111)$	$352(\pm 72)\sum \mathrm{SF} + 615(\pm 87)$	$333(\pm 69)\sum \mathrm{SF} + 612(\pm 89)$
en_core_web_lg	$250(\pm 82)\sum \mathrm{SF} + 710(\pm 108)$	$340(\pm 71)\sum \mathrm{SF} + 627(\pm 88)$	$321(\pm 68)\sum \mathrm{SF} + 621(\pm 89)$
en_core_web_trf	$223(\pm 66)\sum \mathrm{SF} + 734(\pm 91)$	$297(\pm 64)\sum \mathrm{SF} + 682(\pm 79)$	$273(\pm 62)\sum \mathrm{SF} + 684(\pm 82)$
(*b*) Grodner *et al.* dataset			
en_core_web_sm	$87(\pm 22)\mathrm{SF} + 376(\pm 12)$	$68(\pm 31)\mathrm{SF} + 386(\pm 16)$	$108(\pm 27)\mathrm{SF} + 367(\pm 13)$
en_core_web_lg	$74(\pm 20)\mathrm{SF} + 382(\pm 11)$	$70(\pm 29)\mathrm{SF} + 386(\pm 15)$	$90(\pm 27)\mathrm{SF} + 376(\pm 14)$
en_core_web_trf	$65(\pm 17)\mathrm{SF} + 385(\pm 10)$	$60(\pm 26)\mathrm{SF} + 390(\pm 13)$	$75(\pm 26)\mathrm{SF} + 383(\pm 13)$

(c) Predicted garden-path effects for different models

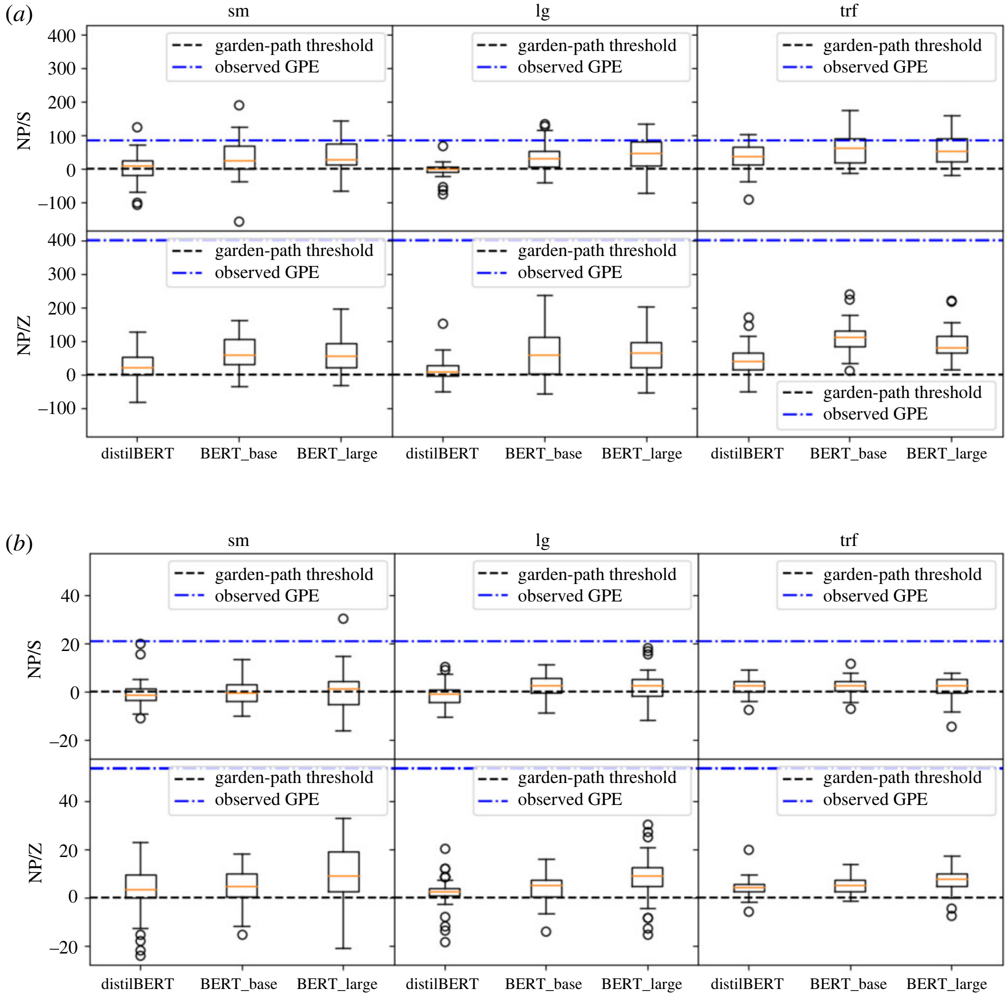


Figure 3. Garden-path effect predictions. The spaCy models are labelled at the top and the BERT models at the bottom. (*a*) Sturt *et al.* dataset. (*b*) Grodner *et al.* dataset.

References

1. Leray J. 1959 Théorie des points fixes: indice total et nombre de Lefschetz. *Bull. Soc. Math. Fr.* **87**, 221–233. (doi:10.24033/bsmf.1519)
2. Grothendieck A. 1957 Sur quelques points d'algèbre homologique, I. *Tohoku Math. J.* **9**, 119–221. (doi:10.2748/tmj/1178244839)
3. Abramsky S, Brandenburger A. 2011 The sheaf-theoretic structure of non-locality and contextuality. *New J. Phys.* **13**, 113036. (doi:10.1088/1367-2630/13/11/113036)
4. Abramsky S, Barbosa RS, Kishida K, Lal R, Mansfield S. 2015 Contextuality, cohomology, and paradox. In *Twenty-Fourth EACSL Annual Conf. on Computer Science Logic*, vol. 41, Berlin, Germany, 6-9 September 2015, pp. 211–228. Leibniz International Proceedings in Informatics. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik.
5. Kishida K. 2023 Godel, Escher, Bell: contextual semantics of logical paradoxes. In *Logic and structure in computer science and beyond, outstanding contributions to logic in honour of Samson Abramsky* (eds *A* Palmigiano, M Sadrzadeh), vol. 25, pp. 531-572. Berlin, Germany: Springer.

6. Wang D, Sadrzadeh M, Abramsky S, Cervantes V. 2021 Analysing ambiguous nouns and verbs with quantum contextuality tools. *J. Cogn. Sci.* **22**, 391–420. (doi:10.17791/jcs.2021.22.3.391)
7. Wang D, Sadrzadeh M, Abramsky S, Cervantes V. 2021 On the quantum-like contextuality of ambiguous phrases. In *Proc. of the 2021 Workshop on Semantic Spaces at the Intersection of NLP, Physics, and Cognitive Science*, Online, 14-18 June 2021, pp. 42–52. Groningen, The Netherlands: Association for Computational Linguistics. (doi:10.48550/arxiv.2107.14589)
8. Wang D, Sadrzadeh M, Abramsky S, Cervantes V. 2021 In search of true contextuality in natural language. Fourth Workshop on Quantum Contextuality and Quantum Mechanics and Beyond (QCQMB). Outstanding Paper Award.
9. Lo KI, Sadrzadeh M, Mansfield S. 2022 A model of anaphoric ambiguities using sheaf theoretic quantum-like contextuality and BERT. In *Proc. End-to-End Compositional Models of Vector-Based Semantics*, NUI Galway, 15–16 August 2022 (eds M Moortgat, G Wijnholds), vol. 366. Electronic Proceedings in Theoretical Computer Science, pp. 23–34. Open Publishing Association. (doi:10.4204/EPTCS.366.5)
10. Lo KI, Sadrzadeh M, Mansfield S. 2023 Generalised winograd schema and its contextuality. In *Proc. of 20th Int. Conf. on Quantum Physics and Logic*, Paris, France, 17-21 July 2023, vol 384, pp. 187–202. Open Publishing Association.
11. Bever T. 1970 The cognitive basis for linguistic structures. In *Cognition and the Development of Language* (ed JR Hayes), pp. 279–362. New York, NY: John Wiley & Sons.
12. Smith NJ, Levy R. 2013 The effect of word predictability on reading time is logarithmic. *Cognition* **128**, 302–319. (doi:10.1016/j.cognition.2013.02.013)
13. Schijndel MV, Linzen T. 2018 Modeling garden path effects without explicit hierarchical syntax. *CogSci.* **2018**, 2603–2608.
14. Dzhafarov EN, Zhang R, Kujala J. 2016 Is there contextuality in behavioural and social systems? *Phil. Trans. R. Soc. A* **374**, 20150099. (doi:10.1098/rsta.2015.0099)
15. Dzhafarov EN, Kujala JV. 2016 Context–content systems of random variables: the contextuality-by-default theory. *J. Math. Psychol.* **74**, 11–33. (doi:10.1016/j.jmp.2016.04.010). Foundations of Probability Theory in Psychology and Beyond.
16. Vallée K, Emeriau PE, Bourdoncle B, Sohbi A, Mansfield S, Markham D. 2023 Corrected bell and non-contextuality inequalities for realistic experiments. (https://arXiv.org/abs/2310.19383)
17. Gogioso S, Pinzani N. 2021 The sheaf-theoretic structure of definite causality. *Electron. Proc. Theor. Comput. Sci.* **343**, 301–324. (doi:10.4204/eptcs.343.13)
18. Gogioso S, Pinzani N. 2023 The geometry of causality. (https://arxiv.org/abs/2303.09017)
19. Abramsky S, Barbosa RS, Searle A. 2023 Combining contextuality and causality: a game semantics approach. (https://arxiv.org/abs/2307.04786)
20. Wang D, Sadrzadeh M. 2022 The causal structure of semantic ambiguities. *Proc. of the 19th Int. Conf. on Quantum Physics and Logic*, vol. 392, Oxford, UK, 27 June - 1 July 2022, pp. 208–220. Open Publishing Association. (https://arxiv.org/abs/2206.06807)
21. Sturt P, Pickering MJ, Crocker MW. 1999 Structural change and reanalysis difficulty in language comprehension. *J. Mem. Lang.* **40**, 136–150. (doi:10.1006/jmla.1998.2606)
22. Grodner D, Gibson E, Argaman V, Babyonyshev M. 2003 Against repair-based reanalysis in sentence comprehension. *J. Psycholinguist. Res.* **32**, 141–166. (doi:10.1023/A:1022496223965)
23. Ehrlich SF, Rayner K. 1981 Contextual effects on word perception and eye movements during reading. *J. Verbal Learn. Verbal Behav.* **20**, 641–655. (doi:10.1016/S0022-5371(81)90220-6)
24. Shannon CE. 1948 A mathematical theory of communication. *Bell Syst. Technical J.* **27**, 379–423. (doi:10.1002/j.1538-7305.1948.tb01338.x)
25. Hale J. 2001 A probabilistic earley parser as a psycholinguistic model. In *Proc. of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies NAACL '01*, Pittsburgh, PA, 2-7 June 2001, pp. 1–8. Pittsburgh, PA: Association for Computational Linguistics. (doi:10.3115/1073336.1073357)
26. van Schijndel M, Linzen T. 2021 Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cogn. Sci.* **45**, e12988. (doi:10.1111/cogs.12988)
27. Huang KJ, Arehalli S, Kugemoto M, Muxica C, Prasad G, Dillon B, Linzen T. 2023 Surprisal does not explain syntactic disambiguation difficulty: evidence from a large-scale benchmark. (doi:10.31234/osf.io/z38u6)

28. Arehalli S, Dillon B, Linzen T. 2022 Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proc. of the 26th Conf. on Computational Natural Language Learning* (*CoNLL*), Abu Dhabi, UAE (Hybrid), 7-8 December 2022, pp. 301–313. Pittsburgh, PA: Association for Computational Linguistics.
29. Kochen S, Specker EP. 1967 The problem of hidden variables in quantum mechanics. *J. Math. Mech.* **17**, 59–87. (doi:10.1512/iumj.1968.17.17004)
30. Einstein A, Podolsky B, Rosen N. 1935 Can quantum-mechanical description of physical reality be considered complete?. *Phys. Rev.* **47**, 777–780. (doi:10.1103/PhysRev.47.777)
31. Bell JS. 1964 On the Einstein Podolsky Rosen paradox. *Phys. Phys. Fizika* **1**, 195–200. (doi:10.1103/PhysicsPhysiqueFizika.1.195)
32. Mansfield S, Kashefi E. 2018 Quantum advantage from sequential-transformation contextuality. *Phys. Rev. Lett.* **121**, 230401. (doi:10.1103/physrevlett.121.230401)
33. Frisson S, Pickering M. 2009 Semantic underspecification in language processing. *Lang. Linguist. Compass* **3**, 111–127. (doi:10.1111/j.1749-818X.2008.00104.x)
34. Pickering M, Frisson S. 2001 Obtaining a figurative interpretation of a word: support for underspecification. *Metaphor Symbol* **16**, 149–171. (doi:10.1207/S15327868MS1603&4_3)
35. Robinson JJ. 1970 Dependency structures and transformational rules. *Language* **46**, 259–285. (doi:10.2307/412278)
36. Vorob'ev NN. 1962 Consistent families of measures and their extensions. *Theory Probab. Appl.* **7**, 147–163. (doi:10.1137/1107014)
37. Barbosa RS. 2014 On monogamy of non-locality and macroscopic averages: examples and preliminary results. *Electron. Proc. Theor. Comput. Sci.* **172**, 36–55. (doi:10.4204/eptcs.172.4)
38. Choi JD, Tetreault J, Stent A. 2015 It depends: dependency parser comparison using a web-based evaluation tool. In *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing* (*Volume 1: Long Papers*), Beijing, China, 26-31 July 2015, pp. 387–396. Pittsburgh, PA: Association for Computational Linguistics.
39. Devlin J, Chang M, Lee K, Toutanova K. 2019 BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186. (doi:10.18653/v1/N19-1423)
40. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. 2017 Attention is all you need. (https://arxiv.org/abs/1706.03762)
41. Wang D. 2023 Empirical models and signalling fractions of garden-path sentences. See https://github.com/wangdaphne/garden-path-SF-dataset.
42. Pickering MJ, Traxler MJ. 1998 Plausibility and recovery from garden paths: an eye-tracking study. *J. Exp. Psychol.: Learn. Mem. Cogn.* **24**, 940–961.
43. Garnsey SM, Pearlmutter NJ, Myers E, Lotocky MA. 1997 The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *J. Mem. Lang.* **37**, 58–93. (doi:10.1006/jmla.1997.2512)