The Semantics Latent in Shannon Information

Alistair M. C. Isaac

ABSTRACT

The lore is that standard information theory provides an analysis of information quantity, but not of information content. I argue this lore is incorrect, and there is an adequate informational semantics latent in standard theory. The roots of this notion of content can be traced to the secret parallel development of an information theory equivalent to Shannon's by Turing at Bletchley Park, and it has been suggested independently in recent work by Skyrms and Bullinaria and Levy. This article explicitly articulates the semantics latent in information theory and defends it as an adequate theory of information content, or natural meaning. I argue that this theory suggests a new perspective on the classic misrepresentation worry for correlation-based semantics.

1 Introduction

The *locus classicus* for information theory is Shannon's ([1948]) 'Mathematical Theory of Communication'. Shannon considered the problem of how much redundancy a communication channel needs to ensure uncertainty about the signal stays below an acceptable threshold. In order to analyse this problem, Shannon modelled the source of the signal as an ergodic Markov process and measured the uncertainty in that process by the weighted average of the log probabilities of each symbol, or 'entropy'. This way of conceptualizing the task of information theory motivated Shannon's ([1948], p. 379) remark that, while such strings of symbols 'Frequently [. . .] have meaning [. . .], [t]hese

doi:10.1093/bjps/axx029

semantic aspects of communication are irrelevant to the engineering problem'. Thus, the lore that Shannon's theory provides no apparatus for analysing information content was born.

Yet Shannon's was only one of two parallel endeavours to mathematically analyse information. A formal apparatus analogous to Shannon's had already been developed independently at Bletchley Park by Turing and colleagues in their daily attempts to crack the Enigma code. While much of the maths was the same (in particular, the appeal to log probabilities as the measure of information, Good [1979]), the goal of Turing's project was radically different, namely to infer from an opaque string of symbols its intended meaning and, more generally, the Enigma machine settings encoding all German messages that day. Thus, whereas Shannon's project was unconcerned with meaning *per se*, Turing's was focused on meaning above all else—not the logician's strict notion of meaning as binary truth conditions, but rather meaning in the sense of significance, or 'weight of evidence', of a signal in favour of one possible interpretation as opposed to others. More recent work suggests that this measure of significance may be transformed into a recognizable semantics.

The goal of this article is to motivate a theory of information content derived entirely from the standard information theory of Shannon and Turing. This semantics is suitable as an analysis of the content of natural signs, or the 'natural meaning' of Grice ([1957]). Intuitively, some events convey information about others, and may be interpreted as signs that these other events obtain; for instance, 'smoke is a sign of fire', and thus from observed smoke, we may safely infer the presence of fire. Grice pointed out that we often speak of this relationship as one of meaning—'those dark clouds mean rain'—yet this natural meaning has different properties from the non-natural, or conventional, meaning familiar from the study of language. The question of whether natural meaning is meaning in a strict or merely metaphorical sense is a vexed one, discussed further in Section 2, which serves to position this project against others in the literature. In brief, the attitude adopted here is that natural and conventional meaning are 'species of a common genus' (Barwise and Perry [1983], p. 16), that the theory of information content on offer is a probabilistic alternative to, but not competitor with, so-called 'semantic' theories of information (STI; for example, Dretske [1981]; Floridi [2004]), and that it constitutes a true semantics in the sense that it exhibits the formal features constitutive of any semantic theory.

The basic semantic model is introduced in Section 3; I claim that information content may be exhaustively represented by a vector of log probability ratios, or 's-vector'. The s-vector encapsulates in a single formal object the complete significance of a signal or event—intuitively, what it 'says about the world'—and in this sense constitutes an analysis of the event's natural meaning. The remainder of the article aims to justify s-vector semantics and elaborate its consequences. Section 4 examines the close connection between meaning and inference, rehearsing the conceptual argument for s-vector semantics offered by Skyrms ([2010]). Section 5 argues that the work of Bullinaria and Levy ([2007], [2012]) empirically validates s-vector semantics by demonstrating it outperforms other correlation-based models of content on semantic tasks. This research suggests a new perspective on the relationship between natural and conventional meaning in language; in brief: words bear natural meaning about other words, and, though it is not equivalent to the conventional meaning they bear about the world, this natural meaning nevertheless determines some of their paradigmatically semantic features. These considerations lead naturally to the final section of the article, which addresses the problem of error for information-based semantics. I argue that (potential) violations of the ergodicity assumption on which information theory is founded suggest a novel route toward the naturalization of misrepresentation.

2 From Correlation to Meaning

Philosophical theories of meaning typically address two types of question: first, what contents should be assigned to a set of meaning-bearing elements; second, in virtue of what do these elements bear the contents they do (Speaks [2016]). The theory offered here assumes an answer to the second question in order to offer an answer to the first. In particular, Shannon information is defined in terms of a probability distribution over events, and thus it is in virtue of patterns in this distribution, in particular statistical correlations, that some events bear content about others. Such correlation-based semantics face several conceptual challenges, and this section briefly considers some of the issues at stake, situating the present project with respect to previous work.

Since at least Dretske ([1981]), philosophers have taken Shannon's admonition that 'meaning [is] irrelevant' in information theory to imply that the mathematics of information requires a supplementary formal system to serve as its semantics. These STIs borrow methods from logic to characterize information content in a manner readily identifiable as propositional. One common strategy, for instance, captures the insight that information reduces uncertainty by treating information content as a set of possible worlds, and information update as changes to the set of worlds available (for example, Dretske [1981]; Floridi [2004]; van Benthem [2011]). A second strategy treats information content as a gappy proposition, filled in by context or background knowledge (for example, situation semantics: Barwise and Perry [1983]; Israel and Perry [1990]). Although they differ in formal specifics, STIs share commitments that contrast helpfully with the view developed here.

Combined with Grice's distinction between natural and non-natural meaning, the STI programme motivates a three-fold categorization of information.

Piccinini and Scarantino ([2011]) helpfully articulate this as a distinction between Shannon ('non-semantic'), natural semantic, and non-natural semantic forms of information. The first obtains whenever the preconditions of Shannon's theory are met, that is, a sequence of events or signals may be modelled by an ergodic Markov process; the second obtains when events are assigned a meaning by an STI, yet the relationship between these events and those they indicate is determined by facts about the world, typically causal or lawlike (Dretske [1981]), or robust correlations within a circumscribed spatiotemporal domain (Millikan [2004]). Semantic non-natural information does not depend on nomic or statistical dependence between signal and signified, but rather on a relationship established by convention, learning, or evolutionary process (Piccinini and Scarantino [2011], Section 4.2).

If, however, there is a semantics latent in Shannon's theory, as argued below, then whenever Shannon's preconditions are satisfied, the events or signals that satisfy them are meaningful. Thus, the distinction between 'natural semantic' information and Shannon information is not best understood as that between meaningful ('semantic') and meaningless information, but rather between information meaningful in one sense (that of STIs), and that meaningful in a different sense (that of s-vector semantics). This perspective agrees with the taxonomy of Piccinini and Scarantino, acknowledging three distinct, progressively more semantically robust types of information, but disagrees with their conclusion that the weakest of these is not meaningful at all. In support of their conclusion, Piccinini and Scarantino rightly point out that the set of elements over which Shannon's theory is defined need not be semantic in the sense of 'stand[ing] for anything' outside of that set (p. 19). However, if the preconditions of Shannon's theory are satisfied, then these elements stand in stable correlation relations with each other, and thus convey meaning about other elements in the set. This 'internal' meaning, meaning in one signal about other signals, was critical for Turing's project at Bletchley Park; furthermore, it seems an appropriate notion of meaning for natural signs, which are themselves merely elements in a set of correlated natural events that convey information about each other.[1]

Nevertheless, Piccinini and Scarantino's presentation highlights two technical challenges for any attempt to ground content in correlation; these

[1] Arguably, all three notions of content are important for decoding an Enigma message. The ultimate target is the conventional meaning of the original German message; in order to uncover it, however, the observed, coded string of symbols is taken to bear natural (STI) meaning about the original message (because their relationship is determined by the lawlike process of Enigma encoding, and thus subject to binary truth conditions). In order to determine the Enigma settings for the day, however, symbols within the coded message must be taken as bearing natural (s-vector) meaning about other symbols, as this internal meaning supports inferences about the underlying correlational structure within the pseudorandom string (cf. Sections 4 and 5).

challenges correspond to two asymmetries in our intuitive understanding of meaning. First, semantic analysis presupposes an asymmetry between signifier and signified: aboutness, reference, representation, and other semantic relations are constitutively directional—'dog' refers to furry, tail-wagging quadrupeds, but those quadrupeds themselves do not likewise refer to 'dog'. Yet typical measures of correlation (the Pearson correlation coefficient; mutual information) are symmetrical; if correlation is to serve as a basis for content, some asymmetrical, directed relationship must be derived from this apparently symmetrical one. Second, and more generally, we typically assign semantic content to specialized objects (words, signals, and so on), not to all possible events. Yet correlations are defined over a homogeneous set of elements, and thus any assignment of content grounded entirely in correlation would seem to assign contents indiscriminately—not only to events typically understood as meaningful, but to all events. As discussed below, the theory presented here exhibits the first asymmetry, between signifier and signified, and may model the second, between those events that bear meaning and those that do not.

A more subtle issue for theories of natural meaning is the question of 'factivity'. Grice argues that it is inconsistent to assert both 'those clouds mean rain' and 'nevertheless, it won't rain'.[2] More generally, he has been interpreted as demonstrating that, if x naturally means y, then if x obtains, y must obtain. Others have taken factivity to be a conceptual condition on the notion of information, that is, x may only bear the information that y if in fact y (Israel and Perry [1990]; Floridi [2007]). However, this view is in tension (on the one hand) with the idea that natural meaning supervenes on correlations, since these are inherently probabilistic, and thus (apparently) x may bear the (Shannon) information that y is probable, and yet y not in fact obtain. It is in tension (on the other) with the project that has motivated much discussion of natural meaning, that of naturalizing meaning *tout court*. If signals bearing non-natural meaning may be tokened in error—I may assert 'McKinley was the twenty-eighth president' when in fact he was the twenty-fifth—yet naturally meaningful signals may not, then it seems that this 'problem of error' poses a significant barrier to any attempt to reduce non-natural meaning to natural meaning.[3] The perspective taken here is that natural meaning may supervene on probabilistic relations without violating factivity; nevertheless, I believe that s-vector semantics sheds new light on the prospects for a

[2] As a conclusion about the concept of natural meaning in ordinary language, Grice's claim is not unassailable (Hazlett [2010]; Isaac [2010], pp. 132–40).

[3] Historically, this problem motivated the shift toward teleological strategies for naturalizing content (Millikan [1984]; Dretske [1988]), although debate continues about the extent to which teleosemantics itself relies on patterns of correlation in the environment (Shea [2007]), provides a satisfactory account of misrepresentation (Fodor [1990]), or, indeed, addresses the problem of naturalizing semantics at all (Godfrey-Smith [2006]).

naturalistic account of signal error. This contentious topic is discussed further in Section 6.

In the face of disagreement about whether vehicles of information are vehicles of meaning, and whether they may be tokened in error or not, by what lights may I claim that the theory on offer here should be understood as properly semantic? I take the constitutive feature of a semantics to be that it assigns a unique, evaluable formal object to each element in a set that characterizes all and only the content conveyed by that element—intuitively, what it 'says about the world'. S-vector semantics provides this for any set of events that satisfies the formal preconditions for Shannon information, just as the formal semantics developed in logic and linguistics provide it for paradigmatically meaningful symbol systems. Some (including many proponents of STIs) have insisted that content must be propositional, but I take this requirement to be essentially vacuous if it is understood as requiring anything stronger than that semantic objects be evaluable (as s-vectors are).[4] Finally, as elaborated below, the theory on offer supports solutions to meaning-requiring tasks, such as determination of semantic categories in a natural language, or effective inference about the true state of the world. I take these features to empirically validate s-vector semantics as an analysis of meaning proper.

3 S-Vector Semantics

Shannon's theory of information presupposes that a sequence of signals or events may be modelled by an ergodic Markov process; this amounts to the claim that their statistical behaviour may be captured by a stable joint probability distribution. Given this joint distribution, we want to assign a unique formal object to each event that characterizes the information that event conveys (what it 'tells us about the world'). This section elaborates the idea that the information conveyed should be identified with the change in information conditional on the event; in the words of Skyrms ([2010], p. 34), how it 'moves the probabilit[ies]'. The formal object that encapsulates this change in information is the vector of log probability ratios, which I call an s-vector.[5] After motivating the idea that log probability ratios characterize the information

[4] This is because there is no consensus metaphysics of propositions that substantively constrains the notion of propositional content as more than just conveying a state of the world (cf. Haugeland's ([1998/1991], p. 191) related discussion of the impotence of possible worlds semantics for distinguishing between types of content. Skyrms ([2010], p. 42) asserts that propositional content is really just a 'special case of the much richer information-theoretic account of content' modelled by s-vector semantics; while Birch ([2014]) disputes this claim, I take it that the real issue in that debate is whether s-vectors are evaluable, and thus may subvene the possibility of misrepresentation.

[5] The 's' in 's-vector' may be taken to stand for 'Shannon', on whose theory it is based, 'Skyrms', who explicitly defends this version of information semantics, or 'semantic', as Bullinaria and Levy call this same construct a 'semantic vector' (Isaac [2010]).

one event carries about another, I introduce the s-vector as the natural generalization of this idea. The section concludes with some basic features of s-vector semantics as an analysis of natural meaning, defending the claim that it exhibits the asymmetry between signifier and signified we intuitively expect from a theory of meaning.

Consider a finite probability space $\langle \Omega, \mathcal{A}, P \rangle$, where Ω is a finite set, $\mathcal{A}$ is an algebra over Ω, and P is a probability distribution over $\mathcal{A}$. An algebra is a family of subsets closed under complement and union, that is, $e \in \mathcal{A}$ implies $e \subseteq \Omega$; if $e \in \mathcal{A}$, then $-e \in \mathcal{A}$ (where $-e = \Omega - e$); and if $e_1, e_2 \in \mathcal{A}$, then $e_1 \vee e_2 \in \mathcal{A}$ (where $e_1 \vee e_2 = e_1 \cup e_2$). It follows that $e_1 \,\&\, e_2 = e_1 \cap e_2$ is also in $\mathcal{A}$, as $e_1 \,\&\, e_2 = -(-e_1 \vee -e_2) \in \mathcal{A}$.

$\mathcal{A}$ is interpreted as the set of possible events; $-e$ is the event incompatible with e; and $e_1 \,\&\, e_2$ is the event of e_1 and e_2 occurring together. The probability distribution P characterizes the correlations between events. To make contact with relevant discussions by Skyrms, Bullinaria and Levy, Good, and Shannon, I'll typically treat P as a summary of long-term relative frequencies. In principle, however, the apparatus developed here is compatible with other philosophical analyses of probability, for instance as propensities or subjective degrees of belief. Shannon's theory takes the existence of a joint probability distribution as a precondition for information, but it is indifferent to the origin or philosophical interpretation of the underlying probabilities.

Given just the probability measure P, we'd like to characterize the information one event in $\mathcal{A}$ conveys about another. First, however, let's consider the measure of information quantity in a single event, call it I; we'd like $I(e)$ to satisfy several intuitive properties.

(1) If $P(e) = 1$, then the quantity of information provided by e is zero, that is, $I(\Omega) = 0$;

(2) All possible events contain positive information, that is, $I(e) \geq 0$ for all $e \in \mathcal{A}$;

(3) An impossible event conveys infinite information, that is, $P(e) = 0$ implies $I(e) = \infty$.

Furthermore, if events e_1 and e_2 are statistically independent, then we'd like the information conveyed by the joint event $e_1 \,\&\, e_2$ to simply add up the information conveyed by the two events separately.

(4) The information in independent events sums, that is, $P(e_1 \,\&\, e_2) = P(e_1)P(e_2)$ implies $I(e_1 \,\&\, e_2) = I(e_1) + I(e_2)$.

A function that satisfies these constraints is the negative log of the probability (independent of choice of base):

$$I(e) = -\log P(e).$$

This function captures our intuitions that the lower the probability of an event, the more information it contains; the certain event contains no information; as an event approaches impossibility, its informational value grows exponentially; and if two unrelated events occur, we gain the complete information from each of them. It turns out that any decreasing function of P that satisfies Condition 4 will also satisfy Conditions 1–3 and be proportional to the negative log; this result confirms the choice of I as the measure of information in an event (Osteyee and Good [1974]). This is the formal notion of information that underlies Shannon's entropy measure, which is just the weighted average of the information quantity of each event in a partition of $\mathcal{A}$, for example, for $e_i \in \Omega$,

$$H = -\sum_i P(e_i) \log P(e_i) = \sum_i P(e_i) I(e_i).$$

Now, for any two events $e_1, e_2 \in \mathcal{A}$, what information does e_1 convey about e_2? We can reconceive this as a quantitative question: How does e_1 change our information about the possibility of e_2? This question was conceived by Turing (as channelled by Good [1950], [1979]) as the question: How does the 'evidence' e_1 affect our assessment of the 'hypothesis' e_2? Turing and Good take this to be the log ratio between the probability of e_2 given e_1 and the prior probability of e_2. The basic idea is that subtracting the information in e_2, given e_1, from the prior information in e_2, measures the change in information about e_2, that is, the information about e_2 conveyed by e_1[6]:

$$I(e_1 : e_2) = I(e_2) - I(e_2|e_1) = -\log P(e_2) + \log P(e_2|e_1) = \log \frac{P(e_2|e_1)}{P(e_2)}$$

This definition has the intuitive features we want in a measure of information conveyed by one event about another:

(1) If $P(e_2|e_1) = P(e_2)$, then e_1 conveys nothing about e_2, and $I(e_1 : e_2) = 0$;

(2) As $P(e_2|e_1)$ grows larger than $P(e_2)$, e_1 conveys more information in favour of e_2 occurring, and $I(e_1 : e_2)$ grows more and more positive;

(3) As $P(e_2|e_1)$ shrinks smaller than $P(e_2)$, e_1 conveys more information against the occurrence of e_2, and $I(e_1 : e_2)$ grows more and more negative.

[6] Strictly speaking, when presenting this definition, Good conceives of e_1 as the hypothesis and e_2 as the evidence; however, the two expressions are equivalent:

$$\frac{P(e_2|e_1)}{P(e_2)} = \frac{P(e_2|e_1)P(e_1)}{P(e_2)P(e_1)} = \frac{P(e_2 \,\&\, e_1)}{P(e_2)P(e_1)} = \frac{P(e_1|e_2)P(e_2)}{P(e_2)P(e_1)} = \frac{P(e_1|e_2)}{P(e_1)}.$$

Here I give the version that conforms with later discussion (cf. Skyrms [2010], p. 35, footnote 4).

Finally, since Shannon's theory makes no assumptions about the relationship between e_1 and e_2 other than the correlation given by the probability distribution P, I characterizes the complete information about e_2 conveyed by e_1.

I measures the information content in one event about another, but what is the total information content of an event? We want a single, unique formal object that captures the information e conveys about all possible events (everything it 'says about the world'). One strategy is simply to collect these separate pieces of information content into a single object that nevertheless keeps them distinct: for instance, a vector. Since $\mathcal{A}$ is finite, we can index it by the natural numbers, and use this enumeration to characterize the full content ν of e with the 's-vector' $\mathrm{v}(e)$:

$$\mathrm{v}(e) = \left\langle \log \frac{P(e_1|e)}{P(e_1)}, \log \frac{P(e_2|e)}{P(e_2)}, \log \frac{P(e_3|e)}{P(e_3)}, \ldots \right\rangle.$$

$\mathrm{v}(e)$ is the basic semantic object implicit in standard information theory, encapsulating the total information content of the event e.

The central claim of s-vector semantics is that e means $\mathrm{v}(e)$ (in the natural, informational sense). At first blush this might seem counterintuitive—a vector of real numbers is simply not the sort of thing that signals or events might mean. But this is to confuse the formal object with our interpretation of it (as information semanticists). To compare, STIs identify the meaning of a signal with a set of 'worlds'; however, from a formal standpoint, this is just a mathematical object, a set of arbitrary abstract elements. The semanticist interprets these elements as possible states of the world. Likewise, while formally an s-vector is just an ordered array of real numbers, the s-vector semanticist interprets these numbers as changes in the probabilities of the events e_1, e_2, e_3, and so on

To illustrate the features of s-vector semantics, consider how it handles a typical example of natural meaning, the claim that smoke is a sign of fire. S-vector semantics presupposes a joint probability distribution over a set of events including both smoke and fire. At the slot in the vector $\mathrm{v}(\mathit{smoke})$ corresponding to $\log P(\mathit{fire}|\mathit{smoke})/P(\mathit{fire})$, there will be a relatively large number, indicating the dramatically increased probability of fire given smoke over the prior probability of fire, that is, that smoke 'means' fire. Yet smoke will mean other things as well: there will also be large values at slots in $\mathrm{v}(\mathit{smoke})$ corresponding to the events danger and low visibility. In contrast, events like thunderstorm may correspond to slots with very low negative numbers, indicating their probability has decreased dramatically. In keeping with the spirit of an informational semantics, the content of $\mathrm{v}(\mathit{smoke})$ will depend upon contingent statistical features of the environment: in a world with constant fire, and only occasional smoke, smoke will not mean fire.

Unlike a naïve approach to deriving content from correlation, which identifies the content of one event directly with the (set of) event(s) with which it is (strongly) correlated, s-vector semantics exhibits the asymmetry we expect from a proper theory of meaning. Events are bearers of information, but they are not themselves meanings. On this view, it is neither events, nor even probabilities of events, but changes in the probabilities of events that are conveyed by (are the content of) a meaningful event. So, while e means $\nu(e)$, that is, a change in the probability distribution over possible ways the world might be, $\nu(e)$ does not 'mean' e, since it is not itself a bearer of meaning, as desired.

The sense in which s-vector semantics is most clearly analogous to our intuitive understanding of meaning, for instance in natural language, is that it supports a graded synonymy relation. If e_1 and e_2 are very close in meaning ('say' very similar things about the state of the world), then $\nu(e_1)$ and $\nu(e_2)$ will be geometrically 'close' by any plausible measure of vector distance. Dark clouds and low pressure both mean rain; correspondingly, their respective s-vectors will fall close together.[7] As e_1 and e_2 approach complete antonymy (convey maximally incompatible states of the world), their corresponding meanings approach $\nu(e_1) = -\nu(e_2)$.[8] Note, however, that antonymy is not the same as negation: $e_1 = -e_2$ does not in general imply that $\nu(e_1) = -\nu(e_2)$. The reason for this is that the value of $\nu(e)$ depends on the correlation between e and other possible events; however, the correlation between e_1 and e_i may not be inversely, or even systematically, related to the correlation between $-e_1$ and e_i. If we find out that the die came up two, we learn with certainty it came up even, but if we learn the die did not come up two, we only learn a little bit about whether it came up even; likewise, red leaves may mean autumn, but no red leaves may not mean much at all about autumn one way or another.

This is a particular instance of a more general feature of s-vector semantics that distinguishes it from typical formal semantics—it is not recursively defined. There is no general relationship between $\nu(e_1)$, $\nu(e_2)$, and $\nu(e_1 \,\&\, e_2)$; this is because the degree of correlation between e_1 and e_2 does not systematically determine their respective correlations with other events, yet these correlations are what determine the relevant s-vectors. This is a straightforward consequence of taking informational content to be determined by a joint probability distribution: since we cannot derive $P(e_i|e_1)$ from $P(e_i)$ and $P(e_1)$, nor $P(e_1 \,\&\, e_2)$ from $P(e_1)$ and $P(e_2)$, we should not expect to be able

[7] This intuition is confirmed by the results of Bullinaria and Levy, as discussed in Section 5.

[8] The negation of a vector is equivalent to the negation of each element within it. Geometrically, $-\nu(e)$ points in the opposite direction from $\nu(e)$.

to derive $\nu(e_1 \& e_2)$ from $\nu(e_1)$ and $\nu(e_2)$. Instead, s-vectors must be defined directly in terms of the joint probability distribution, as this is where the relevant correlations are encoded. If s-vectors are not recursively defined, does that mean they do not constitute a 'semantics'? To conclude as such would constitute a kind of logico-chauvinism, insisting all theories of meaning must conform to one particular style of formal analysis; such chauvinism would not only rule out s-vector semantics, but also other heterodox theories of meaning, for instance holism or contextualism.

4 Meaning for Inference

What of the second asymmetry discussed above? We typically only assign meanings to some events, not all of them. S-vector semantics has the apparatus to model this asymmetry, by treating the set of events over which the joint probability distribution is defined as sorted into two subsets (intuitively: 'signs' and 'signifieds'). The motivation for this sorting comes in part from considerations about how meaningful events or signals are used, namely to support effective inferences about the world. Acknowledging the role of meaning in supporting inference also confirms the formal specifics of the s-vector, which semantically mirror a prominent theory of probabilistic inference, the minimization of Kullback–Leibler divergence.

Skyrms ([2010]) develops an account of the evolution of signalling systems between simple agents, proposing the s-vector as a representation of signal content. His view displays the second semantic asymmetry: signals convey information about other events, but they are not themselves events of semantic interest, that is, they are vehicles, but not topics, of communication. Skyrms takes his analysis of content to be motivated by the use to which signals are put—receivers use them to predict the state of the world, so the correct theory of signal content is the one that specifies the exact inferences a signal supports. Furthermore, this point holds equally for natural signs: we take smoke to be a sign of fire precisely because we can safely infer the existence of fire from smoke, and an adequate analysis of natural meaning should explain how it supports such inferences.[9]

Skyrms's vector is defined by the log ratio between two probability distributions: the prior distribution over states of the world, and the posterior distribution over states of the world, conditional on the signal received. The basic model introduced above may easily be refined to accommodate this interpretation by treating Ω as sorted into two types of primitive event:

[9] The role of natural signs in supporting inferences about the world has been emphasized recently in the literature on probabilistic theories of information (for example, Scarantino and Piccinini [2010], p. 318; Stegmann [2015], Section 4).

states of the world W and signals S, such that $\Omega = W \cup S$. Then the content of each $s_i \in S$ is given by the vector

$$\mathrm{v}(s_i) = \left\langle \log \frac{P(w_1|s_i)}{P(w_1)}, \log \frac{P(w_2|s_i)}{P(w_2)}, \log \frac{P(w_3|s_i)}{P(w_3)}, \ldots \right\rangle,$$

for all $w_j \in W$.

This sorting of events into two types is also a powerful tool for modelling the content of natural signs. Prototypical examples of natural signs '*a* means *b*' are such that *a* is perceptually salient, while *b* is an event of great importance. For instance, smoke is easy to spot from a great distance, or by smell as well as sight, while fire is an event of great importance due to its potential danger and destructive force. Examples like 'these spots are a sign of measles' are even more pronounced: spots are a very easy to see external property, while measles is completely hidden from our regular sensory apparatus, yet a matter of grave concern. We could model this asymmetry by splitting the set of correlated natural events into those that do the natural signing (smoke, spots) and those about which information is conveyed (fire, measles), in complete analogy with Skyrms.

While it is useful to model this apparent asymmetry in paradigmatic cases of natural meaning, it is important to emphasize that the semantic homogeneity of the general s-vector account, which treats every event type as a potential bearer of content, should be considered a feature, not a bug. The information-theoretic perspective on natural meaning is egalitarian about information content: 'The world is full of information' (Skyrms [2010], p. 44). It is only when organisms use the information available in nature, by detecting some events with their perceptual organs, and responding to other events in ways that reflect their importance for survival, that it makes sense to model correlated events as sorted into signs and signifieds. The antecedent presence of these correlations, the simple existence of natural signs, however, is the precondition for this behaviour. The correlations, and thus the information, and thus also the s-vector content, are simply present due to stable facts about the world, independent of any organism detecting this information and using it.[10]

Skyrms's ([2010], p. 42) defence of s-vector semantics emphasizes the close formal connection between the s-vector and Kullback–Leibler (KL)

[10] This point is especially vivid once we recognize that parsing the world into signs and signifieds is both organism and context relative. Many animals use odours as signs for the presence of food, or of a predator, yet these odours are not perceptually salient events for humans. Likewise, while typically one is more likely to infer fire from the presence of smoke, the situation may also be reversed: if fire is visible through the window in a fire-resistant door, I may take the fire as a sign of smoke, and cover my face with my handkerchief before opening it (cf. Cummins *et al.*'s [2006] notion of 'unexploited content').

divergence, a measure for comparing probability distributions.[11] Given the probability measures P and P_e, the KL-divergence of the latter with respect to the former is given by

$$D(P_e||P) = \sum_i P_e(e_i)\log\frac{P_e(e_i)}{P(e_i)}.$$

When $P_e(\cdot) = P(\cdot|e)$, D is just a weighted average over the components of s-vector $\nu(e)$; this averaging erases the particular content conveyed by e in favour of an overall measure of informational divergence of P_e from P.

Why does Skyrms take the formal continuity between s-vectors and KL-divergence to support the claim that the s-vector is the right notion of content for Shannon information? I take it that Skyrms implicitly appeals here to the role of natural meaning in supporting inference. Minimizing KL-divergence is a prominent proposal for how to infer a new probability distribution from a prior plus evidence.[12] If our intuition is that the content of e is just what it tells us about the world, then e's content should be equivalent to whatever we can legitimately infer about the world from it. If the minimal KL-divergence $D(P_e||P)$ identifies the optimal information state to infer from e, then it seems to be the right measure for establishing total content, and the s-vector a legitimate specification of this content, insofar as it unpacks the separate informational relations KL-divergence averages over into a semantically interpretable object. On this view, s-vector semantics inherits conceptual support from any argument that the minimally KL-divergent distribution is exactly what can be inferred (no more, no less) from a piece of evidence.[13]

I take it that Skyrms's considerations provide conceptual support for s-vector semantics. Insofar as minimizing KL-divergence is the optimal theory of information-based inference, and the information content of a signal is equivalent to what it tells us about the world (understood as the inferences it supports), then it appears that the s-vector is the right semantic object for Shannon information. Nevertheless, if our semantic analysis is a purely theoretical exercise, then the s-vector is not the only notion of content that might be derived from a joint probability distribution over events. For instance, Godfrey-Smith ([2012], p. 1292) explores the possibility that one might just take the content of an event to be the posterior probability over states of the world after it occurs, concluding: 'there is probably no need to

[11] Introduced by Kullback and Leibler ([1951]), this measure is sometimes called 'relative entropy' or 'cross-entropy', as it generalizes Shannon's H. Note that D is not symmetric—$D(P_1||P_2) \neq D(P_2||P_1)$—so technically not a distance, hence the term 'divergence'.

[12] In particular, a more general proposal than simple conditionalization; if $P(\cdot|e)$ is not well defined, or if e only loosely constrains posterior probability, minimizing D across distributions that satisfy these constraints determines a unique P_e.

[13] For explicit arguments to this effect see, for instance, (Jaynes [1957]) or (Shore and Johnson [1980]); for relevant surveys see (Domotor *et al.* [1980]) and (Csiszár [2008]).

choose one view, saying that such-and-such is *the* content'. While I agree with the spirit of Godfrey-Smith's remarks, I think there are good reasons to single out the s-vector account once one considers the practical applications of a theory of information content; in particular, the s-vector outperforms other representations of content on semantic tasks, the topic of the next section.

5 Natural Meaning of Conventional Symbols

Skyrms is the first to defend the s-vector as an analysis of content in the philosophical literature; however, essentially the same theory of content appears earlier in (Bullinaria and Levy [2007]). Bullinaria and Levy participate in a research programme that attempts to compute semantic representations from word co-occurrence statistics, testing the validity of these representations on semantic tasks. For instance, given a large corpus, can we derive a representation of the meaning of the word 'hypothesize' from just the relative frequencies of words appearing before (and–or) after it? Can we use this representation to determine whether, say, it is more similar in meaning to 'posit' or 'subjugate'? It turns out that the representation that performs best on semantic tasks like this is the s-vector, a result I take to offer a kind of empirical validation of s-vector semantics.

One might find this result puzzling at first, as words paradigmatically bear conventional meaning, yet I have explicitly offered s-vector semantics as a theory of natural meaning. Whenever a set of items stand in stable probability relations, however, they provide natural meaning about each other that may be characterized by an s-vector. Since words within a corpus do stand in stable probability relations to each other, they convey natural meaning about each other in addition to their conventional meaning. What is surprising about the results discussed here is that the natural meaning words bear about each other in a corpus turns out to be sufficient to solve some semantic tasks typically conceived of as involving conventional meaning. This is good news for naturalistic theories of language acquisition, as it shows that basic semantic relationships may be extracted from a set of words by reinforcement on their correlations. It should be unsurprising to cryptographers, who have relied on the correlation-based information in some parts of a conventionally meaningful text about other parts for thousands of years of code making and breaking. Nevertheless, it is important to emphasize that Bullinaria and Levy and their peers do not pretend that the natural meanings of words extracted from a corpus are equivalent to their full conventional meanings. Rather, they are only able to extract 'some aspects of word meaning' that they posit might 'form a computationally efficient foundation for the learning of semantic representations', perhaps through supervised learning and more elaborate forms of human interaction (Bullinaria and Levy [2007], p. 510).

Since the task of collecting and manipulating word co-occurrence statistics is computationally demanding, research in this field initially proceeded on the basis of *a priori* assumptions about (a) how large or small a contextual window of co-occurring words around the target to consider; (b) how to represent the results of the collected statistics; and (c) what measure of distance between these representations captures degree of semantic 'similarity'. What distinguished Bullinaria and Levy's ([2007]) study at the time was that it treated these as empirical questions.[14] By systematically varying the size of the contextual window that determined their co-occurrence statistics, the manner in which those statistics were represented, and the distance metric between representations, they were able to generate a wide number of different semantic representations, which they then tested on a variety of semantic tasks, such as semantic categorization, syntactic categorization, and synonymy questions from the Test of English as a Foreign Language. Scoring each combination of answers to the three questions on these tests allowed them to empirically determine the optimal semantic representation.

The optimal answer to question (b), the best way to represent co-occurrence statistics for semantic tasks, is as a vector of positive Pointwise Mutual Information (PMI); PMI simpliciter is exactly the same measure as Good's I. Bullinaria and Levy first determined a measure of co-occurrence statistics P, were $P(w)$ is just the number of occurrences of the word w divided by the total number of (token) words in the corpus; the relative frequency of a word w given it appears within the window of co-occurring contextual words around a target word t, $P(w|t)$, is just the number of times w appears with t divided by the total number of appearances of t. Then the PMI 'semantic vector' representing the meaning of a target word t with respect to all potential context words in the corpus, c_i, is given by

$$\left\langle \log \frac{P(c_1|t)}{P(c_1)}, \log \frac{P(c_2|t)}{P(c_2)}, \log \frac{P(c_3|t)}{P(c_3)}, \ldots \right\rangle,$$

that is, identical with $\mathrm{v}(t)$ (pp. 513–14).

Strictly speaking, of all the representations Bullinaria and Levy considered, the PMI vector did worst, while a slight modification of it, the positive PMI vector did best. The positive PMI vector simply replaces all negative-valued components of a PMI vector with zeros. Essentially, PMI simpliciter performed poorly on semantic tasks because very large negative components ensured that some words that should have been judged semantically close were measured as far apart; in the words of Bullinaria and Levy: 'Negative

[14] This potted history is based on discussions in the session 'Lexical Semantics: Bridging the Gap between Semantic Theory and Computational Simulation', at which Bullinaria was an invited speaker, organized by M. Baroni, S. Evert, and A. Lenci at the European Summer School for Logic, Language, and Information, 4–8 August 2008, Hamburg.

values indicate less than the expected number of co-occurrences, which can arise for many reasons, including a poor coverage of the represented words in the corpus' (p. 514). I think this result should still be interpreted on balance as constituting empirical support for the s-vector analysis of information content. Recall that Shannon's theory assumes ergodicity in the information source—this means that in the long term observed statistics will match stable underlying probabilities in the source. The need for positive PMI here, as Bullinaria and Levy acknowledge, is thus due simply to a discrepancy between the assumption of the ideal theory, that observed frequencies match underlying probabilities, and the reality of small data sets. In fact, when tested on an even smaller data set than that initially considered, all Bullinaria and Levy's semantic measures did worse, but the positive PMI outperformed its competitors by an even greater margin.

I take this result, the empirical success of positive PMI vectors on a variety of semantic tasks, to provide a kind of pragmatic validation of the claim that s-vector semantics captures an important notion of information content. Nevertheless, there are significant open questions about the exact implications of this research for a theory of natural meaning. The approach of Bullinaria and Levy is that of the engineer—use whatever achieves results for the task at hand—but an engineering solution does not always conform to our theory-based expectations. In this case, there is some question about the exact significance of the most effective distance measures between semantic vectors. Bullinaria and Levy found that cosine distance (as opposed to, say, Euclidean or city block) between positive PMI vectors produced the best results. In contrast, KL-divergence between probability distributions performed only modestly amongst all measures considered. One might take this result as a mark against a view such as that tentatively advanced by Godfrey-Smith ([2012]), that information content be identified with the posterior probability given the signal, since no distance measure between vectors of posterior probability (including even KL-divergence) performed as well as s-vectors and cosine distance. Nevertheless, the result does seem to undermine the elegant theoretical complementarity between Kullback-Leibler inference and s-vector representation argued in the previous section. What Bullinaria and Levy's results do show is that an information-based semantics may solve real-world semantic tasks, and that optimal performance on such tasks requires a theory very like s-vector semantics.

6 Error and Ergodicity

Do we want a semantics of information on which naturally meaningful events may exhibit error, that is, it is possible that a content-bearing event occur, and yet the actual state of the world not match the content it conveys? As discussed in Section 2, if we accept Grice's argument that natural meaning is factive, then it appears we do not. In contrast, if we think that meaning *tout court* may be naturalized, then perhaps we do want to allow for the possibility of a mismatch between content and world (as long as this mismatch may be explained naturalistically). After a survey of the traditional STI perspective on factivity, error, and natural meaning, I show that (in contrast to the STI view) it is consistent to maintain that natural meaning is purely probabilistic, and yet that it is still, in some sense, 'factive'. Nevertheless, I conclude by considering some ways in which s-vector content might truly fail to match the state of the world, and thus exhibit a naturalistic form of error. The most interesting of these confronts the possibility that Shannon's ergodicity assumption fails, suggesting a new direction for research on the problem of error.

The typical conception of the problem of error for an informational semantics (Dretske [1981]; Fodor [1984]; Godfrey-Smith [1989]), has focused on the causal, nomic, or aetiological relationship supposedly required for information to pass from one event to another. The intuition is that one event cannot convey information about another, if that other event does not in fact occur. Suppose, for instance, that smoke rises from damp, smouldering grass (for example, during the sending of smoke signals), but there is in fact no fire—how could such smoke, then, carry information about fire? If it did bear such information, we could assess the smoke as misrepresenting the state of the world, and make progress on naturalizing error. However, since it does not stand in a causal relationship to any fire, it seems it cannot bear information about fire in the first place. But then the puzzle becomes, how should we distinguish this instance of smoke, which bears no fire information, from other, fire-information-bearing instances of smoke?

As a conceptual problem for information-based semantics, this worry has been extensively discussed for STIs. In general, typical examples of natural signs are not in fact perfectly correlated with the events they naturally 'mean'—sometimes there is smoke without fire. STIs rule out these cases and ensure factivity by stipulating that information is only conveyed naturally under certain circumstances. For instance, Dretske ([1981], p. 65) argues that s may only bear the information content that w if $P(w|s) = 1$ for some nomic reason. For Dretske, in a world where it is possible that smouldering grass produce smoke but not fire, smoke cannot mean fire. Millikan ([2004], Chapter 3) develops a view on which the perfect correlation between events required to ensure factivity obtains within a gerrymandered spatiotemporal region (excluding, say, the smouldering grass), and sign-using organisms succeed in using natural information by 'tracking' these regions. Barwise and Perry ([1983]) appeal to the role of 'constraints' for ensuring factivity of information; in this case the smouldering grass fails the dryness of tinder constraint on the informational relationship between smoke and fire. Sign-users become 'attuned' to these constraints, that is, form habits to act as if they are satisfied, and when environmental constraints change (tinder becomes wet), they may erroneously draw inferences on the assumption of information that is not in fact there (Barwise and Perry [1983], pp. 96–100).

S-vector semantics is not subject to the problem of error construed in this way. To begin with, Shannon does not presuppose that information supervenes on nomic relationships, but on a stable joint probability distribution over events. Since Shannon's theory is blind to whatever underlying causes ensure the stable correlations between events, it does not have the resources to invoke these causes when characterizing information content. Furthermore, s-vector semantics does not equate the information content of *s* directly with some state(s) of the world *w*, but rather with a change in the posterior probability of *w*. Since changes in probability are not themselves factive with regard to states of the world (the probability of rain may increase, and yet it not in fact rain), it seems there is no in principle barrier to the factivity of s-vector semantics. S-vector natural content may be construed as 'factive' in the sense that conveyed changes in probabilities are veridical: if *s* naturally means the probability of *w* increases, and *s* occurs, then the probability of *w* has indeed increased, even if *w* does not actually occur. A smoke event may be caused by smouldering grass, indicate that the probability of fire has dramatically increased, and yet still not be in error. This is because the change in probability the smoke conveys is a fact about the overall statistical co-occurrence of fire and smoke event types, and the absence of a token instance of fire in the case of this particular smoke token does not falsify or contradict that overall pattern of correlation. By treating information as inherently probabilistic, we may avoid the arcane gerrymandering of Dretske and Millikan, while still maintaining the spirit of factivity. This basic insight has been extensively defended in the recent literature on probabilistic information (Scarantino and Piccinini [2010]; Scarantino [2015]; Stegmann [2015]).

Nevertheless, it is also worth considering the possibility that s-vector semantics supports a limited, naturalistic form of information 'error'. If it does, then it may suggest a bridge to span the gap between natural and non-natural forms of meaning, and thereby contribute to an eventual naturalization of misrepresentation. Skyrms ([2010], p. 75) considers the possibility that signals bearing s-vector meaning may misrepresent the world under circumstances where the interests of the signalling agent and the receiver are in conflict. For instance, *Photuris* fireflies 'deceptively' send the mating signals of the *Photinus* genus in order to lure *Photinus* firefly males, which they then eat. This example arguably falls between natural and non-natural meaning—the signalling behaviour of the *Photuris* is the result of reinforcement on correlations, and supervenes on nomic patterns in the environment; nevertheless, it conveys 'misinformation' in the sense that the meaning-bearing event systematically occurs when its reinforced correlate for the receiver (the presence of an actual *Photinus* female) does not. Since the s-vector content of the signal is derived entirely from the joint probability distribution over signals and states of the world, however, the signal 'also increases the probability of a predator'; Skyrms ([2010], pp. 76–7) concludes such 'deceptive' signals convey a kind of 'half-truth'.

While Skyrms describes this example as one of 'misinformation', there is a sense in which the s-vector content of the event continues to veridically match the world, since the changes in probability it conveys do indeed match the actual correlations in the environment: both a *Photinus* female and a predatorial *Photuris* are more likely to be present when the mating signal is sent. In order for an event bearing s-vector content to truly 'misrepresent' the state of the world, the change in probabilities it conveys must fail to match the true pattern of correlations between events. Is such a mismatch possible? We've seen a hint at the answer already in the discussion of Bullinaria and Levy: a signal may misrepresent the world if Shannon's assumption of ergodicity is not satisfied—for instance, if the probabilities that determine its content fail to match the probabilities that obtain when it is tokened.

Bullinaria and Levy believed that their attempt to represent content with vectors of pointwise mutual information was unsuccessful due to a somewhat trivial failure of the ergodicity assumption: their sample set was too small for observed relative frequencies to match 'true' underlying probabilities. Arguably, this is merely an epistemic problem—it is not that PMI semantic vectors misrepresent the correlations between words, but rather that Bullinaria and Levy were unable to determine the true PMI semantic vectors. It is possible, however, that the ergodicity assumption fails for thoroughly metaphysical reasons: underlying probabilities may simply change over time, and thus the observed system of events may fail to be ergodic. Standard information theory, and the definition of the s-vector, presuppose that the probabilistic relations between events are stable. If, conversely, correlations change over time, then the static ratio of probabilities 'meant' by an event in s-vector semantics may fail to match the 'true' probability ratio, and thus that event may 'misrepresent' the state of the world when it occurs.

There is no off-the-shelf metaphysical framework for making sense of standard information theory, and thus s-vector semantics, in a non-ergodic world. Most metaphysics of information is in the STI tradition, focusing on nomic, rather than probabilistic, metaphysical issues. The Skyrms programme investigates signal evolution in a probabilistic world, but typically assumes that world is ergodic. One relevant line of inquiry in this tradition has modelled the repurposing of learned signals to new aspects of the environment (Barrett [2014]), a problem arguably analogous to that of tracking changes in observed correlations. In machine learning, some studies have directly examined strategies for probability matching when causal structure may change and ergodicity is only local (for example, Kummerfeld and Danks [2013]). Nevertheless, a full analysis of this problem, and thus of the problem of error as it applies to s-vector semantics, remains a project for the future.

7 Conclusion

Contrary to popular lore, there is a theory of meaning latent in standard, Shannon information theory: s-vector semantics. Since s-vector semantics rests on the same preconditions as Shannon's theory, it applies whenever a set of events stand in stable correlations with each other. A slogan here might be: where there is information, there is information content. This view contrasts with those that endorse the possibility of 'non-semantic' information in name, but not in spirit. It agrees that there may be true information that does not bear 'semantic content' as defined in the STI tradition, but it claims this information does bear a weak form of content, the content that signals or events in a correlated set convey about each other. S-vector semantics is nevertheless a true semantics, both in the intuitive sense that it assigns a unique semantic object to each event that encapsulates all that it 'says about the world', and in the pragmatic sense that it solves paradigmatically semantic tasks, as demonstrated in the work of Turing and Good, Bullinaria and Levy, and Skyrms.

S-vector semantics is appropriate as a theory of natural meaning, especially the meaning conveyed by natural signs, as this meaning supervenes on stable correlations in the environment. However, conventional symbols may also bear s-vector content when they stand in stable correlations with each other, such as words in a corpus. The s-vector, or natural, meaning borne by conventional symbols is not equivalent to their conventional meaning, it is content they convey about each other, not about the world. It may be surprising to philosophers that this s-vector meaning is nevertheless adequate to sort conventional symbols into semantic and syntactic categories, and to assess relations of synonymy and antonymy, yet this is a feature of the natural meaning in conventional symbol systems that cryptology has relied on for its several thousand year history.

Although s-vector content is inherently probabilistic, it may be viewed through a Gricean lens and interpreted as factive. The insight here is that the change in probabilities conveyed by a signal may be veridical, even when high probability events do not obtain—smoke may convey the information that fire is more likely, even when there is no fire. Nevertheless, one might also wonder if s-vector semantics may subvene a naturalistic account of communication or representation error. I conjecture that some naturalistic 'error', or mismatch between content and world, might emerge in a situation where probabilities are not in fact stable, but change with time. Shannon information, s-vector semantics, and probabilistic theories of inference all typically assume that the world is ergodic, that is, that underlying probabilities remain stable and are reflected in observed, long-run frequencies. Developing a metaphysics for a non-ergodic world, and understanding what information, inference, and natural meaning might amount to in such a world, is a topic for future research.

Acknowledgements

This project has benefitted from the comments of two anonymous reviewers, as well as conversations over the years with Andy Clark, Persi Diaconis, Zoltan Domotor, Carrie Figdor, John Perry, Wolfgang Schwarz, and Brian Skyrms. Some aspects may be found nascent in my doctoral dissertation (Isaac [2010]).

Department of Philosophy
University of Edinburgh
Edinburgh, UK
a.m.c.isaac@ed.ac.uk

References

Barrett, J. A. [2014]: 'Rule-Following and the Evolution of Basic Concepts', *Philosophy of Science*, **81**, pp. 829–39.

Barwise, J. and Perry, J. [1983]: *Situations and Attitudes*, Stanford, CA: CSLI Publications.

Birch, J. [2014]: 'Propositional Content in Signalling Systems', *Philosophical Studies*, **171**, pp. 493–512.

Bullinaria, J. A. and Levy, J. P. [2007]: 'Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study', *Behavior Research Methods*, **39**, pp. 510–26.

Bullinaria, J. A. and Levy, J. P. [2012]: 'Extracting Semantic Representations from Word Co-occurrence Statistics: Stop-Lists, Stemming, and SVD', *Behavior Research Methods*, **44**, pp. 890–907.

Csiszár, I. [2008]: 'Axiomatic Characterizations of Information Measures', *Entropy*, **10**, pp. 261–73.

Cummins, R., Blackmon, J., Byrd, D., Lee, A. and Roth, M. [2006]: 'Unexploited Content', in G. McDonald and D. Papineau (*eds*), *Teleosemantics: New Philosophical Essays*, New York: Oxford University Press, pp. 195–207.

Domotor, Z., Zanotti, M. and Graves, H. [1980]: 'Probability Kinematics', *Synthese*, **44**, pp. 421–42.

Dretske, F. I. [1981]: *Knowledge and the Flow of Information*, Cambridge, MA: MIT Press.

Dretske, F. I. [1988]: *Explaining Behavior: Reasons in a World of Causes*, Cambridge, MA: MIT Press.

Floridi, L. [2004]: 'Outline of a Theory of Strongly Semantic Information', *Minds and Machines*, **14**, pp. 197–222.

Floridi, L. [2007]: 'In Defence of the Veridical Nature of Semantic Information', *European Journal of Analytic Philosophy*, **3**, pp. 31–41.

Fodor, J. A. [1984]: 'Semantics, Wisconsin Style', *Synthese*, **59**, pp. 231–50.

Fodor, J. A. [1990]: 'A Theory of Content, I: The Problem', in *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press, pp. 51–88.

Godfrey-Smith, P. [1989]: 'Misinformation', *Canadian Journal of Philosophy*, **19**, pp. 533–50.

Godfrey-Smith, P. [2006]: 'Mental Representation, Naturalism, and Teleosemantics', in G. MacDonald and D. Papineau (*eds*), *Teleosemantics: New Philosophical Essays*, New York: Oxford University Press, pp. 42–68.

Godfrey-Smith, P. [2012]: 'Book Review: *Signals: Evolution, Learning, and Information*, by Brian Skyrms', *Mind*, **120**, pp. 1288–97.

Good, I. J. [1950]: *Probability and the Weighing of Evidence*, London: Charles Griffin and Co.

Good, I. J. [1979]: 'Studies in the History of Probability and Statistics. XXXVII: A. M. Turing's statistical work in World War II', *Biometrika*, **66**, pp. 393–6.

Grice, H. P. [1957]: 'Meaning', *The Philosophical Review*, **66**, pp. 377–88.

Haugeland, J. [1998/1991]: 'Representational Genera', in *Having Thought*, Cambridge, MA: Harvard University Press, pp. 171–206. Originally published in *Philosophy and Connectionist Theory*, Ramsey, Stich, and Rumelhart (eds.), Lawrence Erlbaum, 1991.

Hazlett, A. [2010]: 'The Myth of Factive Verbs', *Philosophy and Phenomenological Research*, **80**, pp. 497–522.

Isaac, A. M. C. [2010]: 'The Informational Content of Perceptual Experience', Ph.D. thesis, Stanford University, available at <stacks.stanford.edu/file/druid:gb952ys8996/MyThesis-augmented.pdf>.

Israel, D. and Perry, J. [1990]: 'What Is Information?', in P. Hanson (*ed.*), *Information, Language, and Cognition: Vancouver Studies in Cognitive Science*, Vancouver: University of British Columbia Press, pp. 1–19.

Jaynes, E. T. [1957]: 'Information Theory and Statistical Mechanics', *Physical Review*, **106**, pp. 620–30.

Kullback, S. and Leibler, R. A. [1951]: 'On Information and Sufficiency', *The Annals of Mathematical Statistics*, **22**, pp. 79–86.

Kummerfeld, E. and Danks, D. [2013]: 'Tracking Time-varying Graphical Structure', in C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger (*eds*), *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pp. 1205–13, available at < papers.nips.cc/paper/5172-tracking-time-varying-graphical-structure.pdf>.

Millikan, R. G. [1984]: *Language, Thought, and Other Biological Categories*, Cambridge, MA: MIT Press.

Millikan, R. G. [2004]: *Varieties of Meaning*, Cambridge, MA: MIT Press.

Osteyee, D. B. and Good, I. J. [1974]: *Information, Weight of Evidence, the Singularity between Probability Measures, and Signal Detection*, Berlin: Springer-Verlag.

Piccinini, G. and Scarantino, A. [2011]: 'Information Processing, Computation, and Cognition', *Journal of Biological Physics*, **37**, pp. 1–38.

Scarantino, A. [2015]: 'Information as a Probabilistic Difference Maker', *Australasian Journal of Philosophy*, **93**, pp. 419–43.

Scarantino, A. and Piccinini, G. [2010]: 'Information without Truth', *Metaphilosophy*, **41**, pp. 313–30.

Shannon, C. E. [1948]: 'A Mathematical Theory of Communication', *The Bell System Technical Journal*, **27**, pp. 379–423.

Shea, N. [2007]: 'Consumers Need Information: Supplementing Teleosemantics with an Input Condition', *Philosophy and Phenomenological Research*, **75**, pp. 404–35.

Shore, J. E. and Johnson, R. W. [1980]: 'Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy', *IEEE Transactions on Information Theory*, **26**, pp. 26–37.

Skyrms, B. [2010]: *Signals: Evolution, Learning, and Information*, New York: Oxford University Press.

Speaks, J. [2016]: 'Theories of Meaning', in E. N. Zalta (*ed.*), *The Stanford Encyclopedia of Philosophy*, available at <plato.stanford.edu/archives/fall2017/entries/meaning/>.

Stegmann, U. E. [2015]: 'Prospects for Probabilistic Theories of Natural Information', *Erkenntnis*, **80**, pp. 869–93.

van Benthem, J. [2011]: *Logical Dynamics of Information and Interaction*, New York: Cambridge University Press.