SENTENCE COMPREHENSION AS MENTAL SIMULATION: AN INFORMATION-THEORETIC ANALYSIS AND A CONNECTIONIST MODEL

STEFAN L. FRANK

Department of Cognitive, Perceptual and Brain Sciences,
University College London,
London, United Kingdom
E-mail: s.frank@ucl.ac.uk

It has been argued that understanding a sentence comes down to mentally simulating the state-of-affairs described by the sentence. This paper presents a particular formalization of this idea and shows how it gives rise to measures of the amount of syntactic and semantic information conveyed by each word in a sentence. These information measures predict simulated word-processing times in a connectionist model of sentence comprehension.

Keywords: Sentence comprehension; Mental simulation; Word information; Syntax; Semantics; Connectionist modelling; Reading time.

1. Introduction

Generative language models, as defined in the field of Computational Linguistics, capture the occurrence statistics of linguistic forms. If the cognitive system is sensitive to such statistics, language models should be able to account for some aspect of human language processing. Indeed, language models give rise to formal measures of the amount of syntactic information conveyed by each word in a sentence, and it has been suggested that these information measures are related to cognitive processing effort.[1–3] The higher a word's information content, the more 'work' needs to be done to process the word, which would be apparent in, for example, prolonged reading times.

This information-theoretic view of sentence processing may seem at odds with the assumption that our experience with (and knowledge of) the *world*, rather than language, are central to language comprehension. As proponents of this 'embodied linguistics' would claim, the language-comprehension and perceptuomotor systems are deeply intertwined: To understand a sentence is to mentally simulate whatever situation the sentence states to be the case.[4] Consequently, processing difficulty should occur when this situation violates our expectations or experience regarding the way things happen in the world. Take, for example, these two sentences:

(1a) The boys searched for branches with which they went drumming.
(1b) The boys searched for bushes with which they went drumming.

In an ERP experiment,[5] the N400 component on the sentence-final word was found to be larger in (1b) than in (1a), indicating relative comprehension difficulty in (1b). A possible explanation is that mentally simulating the described action is more difficult in (1b) than in (1a), because it makes no sense to try and drum with bushes whereas drumming with branches is possible (albeit somewhat unusual). It is therefore the sentence's meaning in relation to our knowledge of the world, and not our experience with linguistic forms, that is responsible for the N400 effect in this experiment.

The main objective of this paper is to show how the information-theoretic and mental-simulation views on language can be combined by extending the notion of word information to semantics (Sec. 2). Following this, Sec. 3 present a connectionist sentence-comprehension model that treats comprehension as mental simulation, that is, as the construction of a representation of the described situation. Syntactic and semantic word-information measures are defined within this modelling framework. In Sec. 4, the model's word-reading time predictions are compared to the information measures. Indeed, words take longer to process if they convey more information, be it semantic or syntactic. Sec. 5 discusses the implication of these findings for theories of language processing and acquisition, and Sec. 6 concludes.

2. Measuring word information

Formal measures of word information have only been proposed with respect to generative language models, in particular Probabilistic Context-Free Grammars.[1–3,6–10] By definition, a generative language model defines a probability distribution over sentences. As the words of a sentence are processed one at a time, this distribution changes. It is this fluctuation of sentence probabilities that gives rise to measures of word information. Since these sentence probabilities depend only on the statistics of linguistic *forms*, the information measures will be referred to as *syntactic*.

Alternatively, word information content can be defined with respect to the probabilities of different states of the world. When a sentence is processed incrementally, each word changes the probabilities of the states-of-affairs being described. Again, this leads to formalizations of word information. These information measures depend on knowledge of the *meaning* of language and will therefore be referred to as *semantic*.

Syntax and semantics are different sources of information. An orthogonal distinction is between different *measures* of information. Two such measures will be used here: *surprisal* and *entropy reduction*. As explained in more detail below, surprisal is a measure of the unexpectedness of a word's occurrence and entropy reduction quantifies the extent to which a word reduces the uncertainty about the upcoming material.

The remainder of this section explains and formalizes four types of information: syntactic surprisal, syntactic entropy reduction, semantic surprisal, and semantic entropy reduction.

2.1. *Syntactic information measures*

2.1.1. *Language model*

Let $\mathcal{S}$ denote the (possibly infinite) set of all complete sentences. An n-word sentence is a string of words $w_1, \ldots, w_n$, denoted $w_{1\ldots n}$ for short. A generative language model defines a probability distribution over $\mathcal{S}$, such that $P(w_{1\ldots n}) = 0$ if $w_{1\ldots n} \in \mathcal{S}$ is ungrammatical.

As a sentence is processed one word at a time, the sentence probabilities fluctuate. Let $P(w_{1\ldots n}|w_{1\ldots i})$ be the probability of sentence $w_{1\ldots n}$ given that the first i words (i.e., the string $w_{1\ldots i}$) have been seen so far. If $w_{1\ldots i}$ does not match the first i words of $w_{1\ldots n}$ (or if $n < i$) then $P(w_{1\ldots n}|w_{1\ldots i}) = 0$. Otherwise, $w_{1\ldots n}$ equals $w_{1\ldots i,i+1\ldots n}$, and, by definition of conditional probability,

$$P(w_{1\ldots n}|w_{1\ldots i}) = \frac{P(w_{1\ldots i,i+1\ldots n})}{P(w_{1\ldots i})}.$$

Here, the probability of the incomplete sentence $w_{1\ldots i}$ is the total probability of all sentences that begin with $w_{1\ldots i}$, that is,

$$P(w_{1\ldots i}) = \sum_{w_{1\ldots i,i+1\ldots n} \in \mathcal{S}} P(w_{1\ldots i,i+1\ldots n}).$$

2.1.2. *Syntactic surprisal*

Linguistic expressions differ strongly in their occurrence frequencies. For one, idiomatic expressions are often more frequent than similar, nonidiomatic sentences. As a result, the word ‘dogs’ is more expected in the context of (2a) than in (2b):

(2a) It is raining cats and dogs.
(2b) She is training cats and dogs.

Surprisal is a formal measure of the extent to which a word occurs unexpectedly. The syntactic surprisal of word w_{i+1} given the sentence-so-far $w_{1\ldots i}$ is defined as the negative logarithm of the word's probability according to the language model:

$$\begin{aligned} s_{\text{syn}}(i+1) &= -\log P(w_{i+1}|w_{1\ldots i}) \qquad (1) \\ &= \log P(w_{1\ldots i}) - \log P(w_{1\ldots i+1}). \end{aligned}$$

It has been suggested that syntactic surprisal is indicative of cognitive processing effort and should therefore be predictive of word-reading time.[1,3] Indeed, reading times have repeatedly been shown to correlate positively with surprisal as estimated by a range of language models.[7–9,11,12]

2.1.3. *Syntactic entropy reduction*

Before a sentence has been fully processed, there may exist uncertainty about the upcoming words. This uncertainty usually (but not necessarily) decreases with each word that comes in. For example, after processing ‘It is raining’ it is uncertain if the sentence is over, if a connective (like ‘and’) will follow, or if the verb is used in the less frequent transitive sense, as in sentence (2a). Hence, there is quite some uncertainty about what will come next. Presumably, the amount of uncertainty is more or less the same after ‘She is training’. Now assume that the next word turns out to be ‘cats’. In (2a), the occurrence of ‘cats’ make it almost certain that the rest of the sentence will be ‘and dogs’, that is, uncertainty is reduced to nearly zero. In (2b), on the other hard, the occurrence of ‘cats’ is not very informative about what the next words will be, so much uncertainty remains.

Syntactic entropy is a formal measure of the uncertainty about the rest of the sentence. It equals the entropy of the probability distribution over all sentences in $\mathcal{S}$, which is defined as

$$H_{\text{syn}} = -\sum_{w_{1\ldots n}\in\mathcal{S}} P(w_{1\ldots n}) \log P(w_{1\ldots n}).$$

When the sentence's first i words have been processed, and the sentence probabilities have changed accordingly, the entropy is

$$H_{\text{syn}}(i) = -\sum_{w_{1\dots n}\in\mathcal{S}} P(w_{1\dots n}|w_{1\dots i}) \log P(w_{1\dots n}|w_{1\dots i}). \tag{2}$$

Note that entropy is minimal when all probabilities are zero[a] except for a single sentence $w_{1\dots n}$, which must then have a probability of one. In that case, there is certainty about the upcoming words. In contrast, uncertainty (and entropy) is maximal when all sentences have the same probability.

Processing word w_{i+1} reduces the entropy by

$$\Delta H_{\text{syn}}(i+1) = H_{\text{syn}}(i) - H_{\text{syn}}(i+1).$$

This reduction in syntactic entropy due to processing a word has been argued to be a cognitively relevant measure of the amount of information conveyed[2,6] and has recently been shown to predict word-reading times independently of syntactic surprisal.[13,14]

2.2. *Semantic information measures*

2.2.1. *World model*

When defining the syntactic information measures above, the sentence-processing system's task was taken to be the identification of the incoming sentence. Likewise, the task of sentence *comprehension* is to identify the state-of-affairs that is asserted by the sentence. This requires knowledge of the many possible states of the world and their probabilities.

Let $\text{sit}(w_{1\dots n})$ denote the situation described by sentence $w_{1\dots n}$ and $P(\text{sit}(w_{1\dots n}))$ the probability of that situation. Situations can be combined using boolean operators. So, if p and q denote two situations then $\neg p$, $p \wedge q$ and $p \vee q$ are also situations. For example, if $p = \text{sit}(\textit{it is raining})$ and $q = \text{sit}(\textit{it is daytime})$ then $\neg p = \text{sit}(\textit{it is not raining})$ and $p \wedge q = \text{sit}(\textit{it is raining and it is daytime})$.

The situation described by an incomplete sentence, $\text{sit}(w_{1\dots i})$, is defined as the disjunction of all situations described by sentences that start with $w_{1\dots i}$. So, $\text{sit}(\textit{it is}) = \text{sit}(\textit{it is raining}) \vee \text{sit}(\textit{it is daytime}) \vee \text{sit}(\textit{it is freezing}) \vee \dots$.[b]

[a] Strictly speaking, entropy is not defined when there is a zero probability because $\log(0)$ does not exist. However, when p goes to zero, $p\log(p) = 0$ in the limit. Therefore, we can take $0\log(0)$ to equal zero.

[b] This definition assumes that the sentence does not contain a negation. Without this assumption, $\text{sit}(\textit{it is}) = \text{sit}(\textit{it is raining}) \vee \text{sit}(\textit{it is not raining}) \vee \dots$, which would be

2.2.2. *Semantic surprisal*

Situations and events in the world differ in their (perceived) likelihood of occurrence. For example, according to our knowledge of the academia, the situation described in (3a) should be more probable than the one in (3b):

(3a) The brilliant paper was immediately accepted.
(3b) The terrible paper was immediately accepted.

Consequently, the word ‘accepted’ is more expected in (3a) than in (3b). Semantic surprisal quantifies the extent to which the incoming word led to the assertion of a situation that is unlikely to occur, given what was already learned from the sentence so far.

A sentence up to word w_i describes the situation $\mathrm{sit}(w_{1\ldots i})$. The next word, w_{i+1}, changes the situation to $\mathrm{sit}(w_{1\ldots i+1})$. The corresponding change in the situations' probabilities gives rise to a definition of the semantic surprisal of w_{i+1}, analogous to syntactic surprisal of Eq. 1:

$$\begin{aligned} s_{\mathrm{sem}}(i+1) &= -\log P(\mathrm{sit}(w_{1\ldots i+1})|\mathrm{sit}(w_{1\ldots i})) \\ &= \log P(\mathrm{sit}(w_{1\ldots i})) - \log P(\mathrm{sit}(w_{1\ldots i+1}))). \end{aligned} \tag{3}$$

Note that semantic surprisal is independent of the sentence probabilities. Only the probability of the described situation is relevant.

2.2.3. *Semantic entropy reduction*

Brilliant papers get accepted whereas the fate of a mediocre paper is unsure. Therefore, the uncertainty about the situation being communicated is higher in sentence fragment (4a) than in (4b).

(4a) The mediocre paper was immediately —
(4b) The brilliant paper was immediately —

When a sentence is complete (and unambiguous) there is no more uncertainty. Therefore, whether the next word of (4a) and (4b) turns out to be ‘accepted’ or ‘rejected’, it reduces uncertainty more strongly, and therefore conveys more semantic information, in (4a) than in (4b).

uninformative so sentence interpretation could not start until the sentence is over and it is sure the speaker will not suddenly exclaim ‘*Not!*’. In line with this assumption, experimental evidence[15] has indicated that negations are ignored initially and applied only after the negated statement has been mentally represented.

Semantic entropy quantifies the uncertainty about the described situation. It is not as easy to define as syntactic entropy; The main problem here is that the total probability over all situations is larger than 1 because situations are not mutually exclusive. For example, one situation may be that it is raining and another that it is daytime, and these can obviously occur simultaneously. As a result, the collection of situation probabilities does not form a probability distribution. Moreover, situations are not discrete entities like sentences: There seems to be uncountably infinite situations, whereas the number of sentences may be infinite but is at least countable. These two problems are circumvented here by only taking into account those situations that can be described by some sentence, and by normalizing probabilities to sum to 1:

$$P_{\text{norm}}(\text{sit}(w_{1\dots n})) = \frac{P(\text{sit}(w_{1\dots n}))}{\sum_{v_{1\dots n}\in\mathcal{S}} P(\text{sit}(v_{1\dots n}))}.$$

By analogy with syntactic entropy (Eq. 2), the semantic entropy after processing $w_{1\dots i}$ is

$$H_{\text{sem}}(i) = -\sum_{w_{1\dots n}\in\mathcal{S}} P_{\text{norm}}(\text{sit}(w_{1\dots n})|\text{sit}(w_{1\dots i}))\log P_{\text{norm}}(\text{sit}(w_{1\dots n})|\text{sit}(w_{1\dots i})), \tag{4}$$

making the semantic entropy reduction due to word w_{i+1}:

$$\Delta H_{\text{sem}}(i+1) = H_{\text{sem}}(i) - H_{\text{sem}}(i+1).$$

3. The sentence-comprehension model

This section presents a recent connectionist model of sentence comprehension[16] that implements language understanding as mental simulation. The model treats word-by-word processing of a sentence $w_{1\dots n}$ as the incremental construction of a representation of the described situation $\text{sit}(w_{1\dots n})$. As explained in detail below, probabilities of situations follow directly from the representations, making this framework ideally suited for incorporating semantic word-information measures and for studying their effect on sentence processing. A simple extension to the model makes it possible to obtain word-processing times, which are compared to word information.

3.1. *Microworld*

According to the mental-simulation view of language comprehension, understanding a sentence requires real-world knowledge and experience. To make world knowledge manageable for the model, it is restricted here to a small 'microworld'. The rest of this section presents the microworld and explains how states of the microworld are represented in the model. However, many details will be skipped since they can be found elsewhere.[16]

3.1.1. *Situations*

The microworld has only three inhabitants (sophia, heidi, and charlie) and four locations (bedroom, bathroom, street and playground. There also exist games and toys, like chess, hide&seek, ball, and puzzle. All in all, 44 different atomic situations can occur in the world. Examples are play(heidi,chess), win(sophia), and place(charlie,playground), which, respectively, refer to heidi playing chess, sophia winning, and charlie being in the playground. Atomic situations can be combined using the boolean operators of negation, conjunction, and disjunction, creating more complex situations such as play(heidi,chess) $\wedge$ play(charlie,chess) $\wedge\neg$lose(heidi), which is the case when heidi does not lose a game of chess to charlie.

Some situations are more likely to occur than others. To name a few, heidi tends to win at hide&seek, sophia usually loses at chess, and charlie is most often in a different place than the girls. There are also hard constraints on possible situations, for instance, each of the three protagonists is always in exactly one place, the ball is only played with in outside locations, and someone has to play some game in order to win or lose.

3.1.2. *Representation*

Following a scheme originally developed for a model of story comprehension,[17] microworld situations are represented by vectors in a high-dimensional 'situation space'. A crucial property of these representations is that they are analogical, in the sense that relations among the vectors mirror relations among the represented situations. The same has been argued to hold for mental representations.[18]

The vectors for atomic situations follow from a large number of examples of microworld states. Each example shows which atomic situations are the case, and which are not, at one moment in microworld time. The probabilistic constraints on co-occurrences of situations, which are apparent from the examples, are extracted by a self-organizing network. Its output consists of one 150-dimensional vector $\mu(p) = (\mu_1(p), \ldots, \mu_{150}(p)) \in [0,1]^{150}$ for each atomic situation p. These situation vectors are such that approximate microworld probabilities follow directly. Probability estimates from situation vectors, called 'belief values', are denoted by the symbol τ:

$$\tau(p) = \frac{1}{150}\sum_j \mu_j(p) \qquad \approx P(p) \tag{5}$$

$$\tau(p \wedge q) = \frac{1}{150}\sum_j \mu_j(p)\mu_j(q) \approx P(p \wedge q). \tag{6}$$

From Eq. 5 and the fact that $P(\neg p) = 1 - P(p)$, it follows that $\mu(\neg p) = 1 - \mu(p)$. From Eq. 6 it follows that $\mu_j(p \wedge q) = \mu_j(p)\mu_j(q)$. By making use of the fact that $p \vee q \equiv \neg(\neg p \wedge \neg q)$, the vector representing the disjunction $p \vee q$ can also be constructed. In short, any complex situation can be represented by combining the 44 atomic situation vectors, and a vector's average element value is an estimate of the probability of the represented situation.

Another interesting and useful property of situation vectors is that probabilistic inference is performed by the representations themselves. The probability of any (atomic or complex) situation p given that q is the case, can be observed directly from $\mu(p)$ and $\mu(q)$:

$$P(p|q) = \frac{P(p \wedge q)}{P(q)} \approx \frac{\sum_j \mu_j(p)\mu_j(q)}{\sum_j \mu_j(q)} = \frac{\tau(p \wedge q)}{\tau(q)} = \tau(p|q).$$

This means that a representation of the fact that p is also a representation of anything that is implied by p. For example, if sophia is playing with the ball she cannot be in the bedroom, and indeed τ(place(sophia,bedroom)|play(sophia,ball))≈ 0. This property of direct inference, which distinguishes analogical from symbolic representations, is also present in mental representations: In one experiment,[19] participants who were told that *the pencil is in the cup* automatically and unconsciously formed a visual representation in which the pencil is in a vertical position, as opposed to when they heard that *the pencil is in the drawer*.

3.2. *Microlanguage*

Microworld situations are described by microlanguage sentences. Again, full details are published elsewhere[16] so they will not be presented here. The language has a vocabulary of 40 words, including (proper) nouns like *heidi*, *girl*, *playground*, and *chess*; verbs such as *beats*, *is*, and *played*; adverbs (*inside*, *outside*); and prepositions (*with*, *at*, *in*). These words can be combined to form 13,556 different sentences, each unambiguously referring to one (atomic or complex) microworld situation. A few examples are presented in Table 1.

Table 1. Examples of microlanguage sentences and corresponding situations. c = charlie; h = heidi; s = sophia.

Sentence	Situation
charlie plays chess	play(c, chess)
chess is played by charlie	play(c, chess)
girl plays chess	play(h, chess) $\vee$ play(s, chess)
sophia plays with ball in playground	play(s, ball) $\wedge$ place(s, playground)
chess is lost by heidi	lose(h) $\wedge$ play(h, chess)
charlie wins outside	win(c) $\wedge$ (place(c, street) $\vee$ place(c, playground))
sophia beats charlie at hide-and-seek	win(s) $\wedge$ lose(c) $\wedge$ play(s, hide&seek)

3.3. *The model*

3.3.1. *Architecture*

The sentence-comprehension model is a Simple Recurrent Network (SRN)[20] that takes as input a microlanguage sentence, one word at a time, and is trained to give as output the vector representing the corresponding situation. As shown in Fig. 1, the network has a standard three-layer architecture, with 40 input units (one for each word in the microlanguage), 120 hidden (i.e., recurrent) units, and 150 output units (corresponding to the 150 dimensions of situation space).

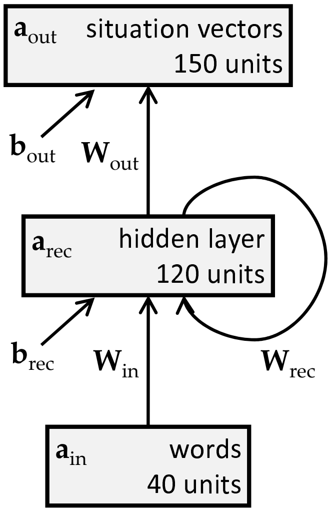
$\mathbf{a}_{\text{out}}$ situation vectors 150 units
$\mathbf{b}_{\text{out}}$
$\mathbf{W}_{\text{out}}$
$\mathbf{a}_{\text{rec}}$ hidden layer 120 units
$\mathbf{b}_{\text{rec}}$
$\mathbf{W}_{\text{in}}$
$\mathbf{W}_{\text{rec}}$
$\mathbf{a}_{\text{in}}$ words 40 units

Fig. 1. Architecture of the sentence-comprehension network.

To process the word occurring at sentence position $i+1$, the common SRN equations are:

$$\begin{aligned}\mathbf{a}_{\text{rec}}(i+1) &= f(\mathbf{W}_{\text{rec}}\mathbf{a}_{\text{rec}}(i) + \mathbf{W}_{\text{in}}\mathbf{a}_{\text{in}}(i+1) + \mathbf{b}_{\text{rec}}) \\ \mathbf{a}_{\text{out}}(i+1) &= f(\mathbf{W}_{\text{out}}\mathbf{a}_{\text{rec}}(i+1) + \mathbf{b}_{\text{out}}),\end{aligned} \tag{7}$$

where $\mathbf{W}$ are connection weight matrices; $\mathbf{b}$ are bias vectors; $f(\mathbf{x})$ is the logistic function; and $\mathbf{a}_{\text{in}}(i+1)$ is the input vector that forms a localist encoding of word w_{i+1}. The sentence-comprehension model follows these equations except that, after training, $\mathbf{a}_{\text{out}}(i+1)$ is computed differently, as explained in Sec. 3.3.3 below. When a sentence that describes microworld situation p has been processed, the output vector $\mathbf{a}_{\text{out}}$ ideally equals p's vector representation $\mu(p)$.

3.3.2. *Network training*

Training examples were randomly sampled from all 13,556 microlanguage sentences, with shorter sentences having a larger sampling probability. Each sampled sentence was presented to the network, one word at a time. At each word, the target output was the vector representing the situation described by the complete sentence, and network weights and biases were updated accordingly using the standard backpropagation algorithm. Training stopped when the squared error between targets $\mu(\text{sit}(w_{1\ldots n}))$ and actual outputs $\mathbf{a}_{\text{out}}(n)$, over all microlanguage sentences $w_{1\ldots n}$, was down to 2% of the pre-training error. It took approximately 1.4 million training sentences to reach that criterion.

3.3.3. *Obtaining processing times*

The original model, presented so far, cannot predict reading times because processing a word always takes one sweep of activation through the network: There is no notion of processing over time. To extract reading-time predictions from the network, processing time must be made to vary over words. This is accomplished by turning the trained network's output vector update (Eq. 7) into a dynamical process. Simply stated, processing the word at $i+1$ involves a change from $\mathbf{a}_{\text{out}}(i)$ to $\mathbf{a}_{\text{out}}(i+1)$ over continuous time t rather than instantaneously. This process follows a simple differential equation:

$$\frac{\mathrm{d}\mathbf{a}_{\text{out}}}{\mathrm{d}t} = \mathbf{a}_{\text{out}} - f(\mathbf{W}_{\text{out}}\mathbf{a}_{\text{rec}}(i+1) + \mathbf{b}_{\text{out}}), \tag{8}$$

where the initial value of $\mathbf{a}_{\text{out}}$ equals $\mathbf{a}_{\text{out}}(i)$, and $\mathbf{a}_{\text{out}}(0)$ is set to the unit vector, which conveys no information about the state of the microworld because $\tau(p|\mathbf{1}) = \tau(p)$ for any p.

According to Eq. 8, the output vector $\mathbf{a}_{\text{out}}$ moves over processing time towards $f(\mathbf{W}_{\text{out}}\mathbf{a}_{\text{rec}}(i+1) + \mathbf{b}_{\text{out}})$ which equals $\mathbf{a}_{\text{out}}(i+1)$ of Eq. 7. The processes converges when $\mathbf{a}_{\text{out}}$ no longer changes, that is, when $\mathrm{d}\mathbf{a}_{\text{out}}/\mathrm{d}t = 0$. This is only the case when $\mathbf{a}_{\text{out}} = \mathbf{a}_{\text{out}}(i{+}1)$. Hence, after convergence, the output vector equals the output of the standard SRN (see Eq. 7). However, convergence is asymptotic so $\mathrm{d}\mathbf{a}_{\text{out}}/\mathrm{d}t$ never quite reaches 0. For this reason, the process is halted when the rate of change in $\mathbf{a}_{\text{out}}$ drops below a certain threshold:

$$|\mathrm{d}\mathbf{a}_{\text{out}}/\mathrm{d}t| < \max\{0.1 \times |\mathbf{a}_{\text{out}}|, 10^{-8}\}, \tag{9}$$

where $|\mathbf{x}|$ denotes the euclidean length of vector $\mathbf{x}$. So, word processing stops when the amount of change in $\mathbf{a}_{\text{out}}$ is less than 10% of the length of $\mathbf{a}_{\text{out}}$ itself, or smaller than 10^{-8}, whatever comes first. The amount of time t required to reach the stopping criterion of Eq. 9 is the simulated reading time on word w_{i+1}.

3.4. *Measures of word information*

The model's reading-time predictions are compared to measures of word information. The current framework allows for definitions of both syntactic and semantic information, both in terms of surprisal and in terms of entropy reduction.

3.4.1. *Syntactic information*

The probabilities $P(w_{i+1}|w_{1\ldots i})$ and $P(w_{1\ldots n}|w_{1\ldots i})$, from the definitions of syntactic surprisal (Eq. 1) and entropy (Eq. 2), are estimated directly from the frequencies with which sentences were sampled for training. This gives

$$\begin{aligned} s_{\text{syn}}(i+1) &= \log(\text{freq}(w_{1\ldots i})) - \log(\text{freq}(w_{1\ldots i+1})) \\ H_{\text{syn}}(i) &= -\sum_{w_{1\ldots i,i+1\ldots n}\in\mathcal{S}} \frac{\text{freq}(w_{1\ldots i,i+1\ldots n})}{\text{freq}(w_{1\ldots i})} \log \frac{\text{freq}(w_{1\ldots i,i+1\ldots n})}{\text{freq}(w_{1\ldots i})}, \end{aligned}$$

where $\text{freq}(w_{1\ldots n})$ is the number of times sentence $w_{1\ldots n}$ was sampled for training, and $\text{freq}(w_{1\ldots i})$ is the total sampling frequency of sentences that begin with $w_{1\ldots i}$.

3.4.2. *Semantic information*

The computation of semantic surprisal (see Eq. 3) requires an estimate of $P(\mathrm{sit}(w_{1\ldots i}))$, the probability of situations described by sentences that start with $w_{1\ldots i}$. Since the set of sentence/semantics-pairs in the microlanguage is known, determining $\mathrm{sit}(w_{1\ldots i})$ is straightforward: It is the disjunction of $\mathrm{sit}(w_{1\ldots n})$ over all sentences $w_{1\ldots n}$ that start with $w_{1\ldots i}$. The probability of $\mathrm{sit}(w_{1\ldots i})$ is estimated by its belief value, which follows from its vector representation:

$$P(\mathrm{sit}(w_{1\ldots i})) \approx \tau(\mathrm{sit}(w_{1\ldots i})) = \frac{1}{150}\sum_j \mu_j(\mathrm{sit}(w_{1\ldots i})).$$

To obtain measures of semantic entropy (Eq. 4), estimates of $P_{\mathrm{norm}}(\mathrm{sit}(w_{1\ldots n})|\mathrm{sit}(w_{1\ldots i}))$ are needed. Let Z denote the set of microworld situations that can be described by some microlanguage sentence. The belief value for each such situation $z \in Z$ is $\tau(z)$. When the sentence-so-far $w_{1\ldots i}$ has been processed, the belief values are $\tau(z|\mathrm{sit}(w_{1\ldots i}))$. These belief values are made to sum to 1, turning them into a proper probability distribution:

$$\begin{aligned}P_{\mathrm{norm}}(\mathrm{sit}(w_{1\ldots n})|\mathrm{sit}(w_{1\ldots i})) &\approx \tau_{\mathrm{norm}}(\mathrm{sit}(w_{1\ldots n})|\mathrm{sit}(w_{1\ldots i}))\\ &= \frac{\tau(\mathrm{sit}(w_{1\ldots n})|\mathrm{sit}(w_{1\ldots i}))}{\sum_{z\in Z}\tau(z|\mathrm{sit}(w_{1\ldots i}))}.\end{aligned}$$

4. Experiments and Results

Everything is now in place to investigate the relation between the four word-information measures and the model's word-processing times. First, for each of the 84,321 word tokens in the 13,556 microlanguage sentences, surprisal and entropy reduction were computed with respect to both syntax and semantics. Next, the trained network received all sentences and processed them one word at a time, yielding a processing time for each word token.

Although all sentences are different, many are identical up to a certain word. The information measures and processing times are also identical up to that word, so many data points occur multiple times. All these copies were removed from the analysis, leaving a total of 15,873 data points.

The predictive value of the four information measures is determined by regressing the word-processing times on these measures. In addition, the position of the word in the sentence is included as a predictor. The pairwise correlations between these five predictors are shown in Table 2.

Table 3 presents the result of a stepwise regression analysis in which the predictor with the highest additional explanatory value was added at each

Table 2. Matrix of correlation coefficients between predictors.

	ΔH_{syn}	s_{sem}	ΔH_{sem}	Position
s_{syn}	.95	.09	.21	$-.03$
ΔH_{syn}		.08	.21	$-.07$
s_{sem}			.42	.28
ΔH_{sem}				.13

step. The table shows both the predictor's coefficient and the fraction of variance explained (R^2) over and above what is accounted for by the predictors already in the model. Each addition resulted in a highly significant ($p < 10^{-8}$) increase in regression model fit.

Table 3. Regression analysis results. Factors are in order of inclusion in the regression model.

Predictor	Coefficient	R^2
s_{sem}	0.04	.310
ΔH_{sem}	0.64	.082
s_{syn}	0.12	.026
Position	0.08	.011
ΔH_{syn}	0.20	.001

By far the largest contribution to explained variance (31.0%) comes from semantic surprisal: A word takes longer to process if it makes the sentence describe a less likely situation. Semantic entropy reduction comprises the second largest contribution (8.2%): A word takes longer to process if it more strongly reduces the uncertainty about the situation being described. Syntactic information measures explain little (2.8% in total), but significant, additional variance in word-processing time. The effect of syntactic entropy reduction is particularly weak (just over 0.1%) but considering the high correlation with syntactic surprisal[c] (see Table 2) the additional variance explained by syntactic entropy reduction could only be very small. In total, 43.01% of variance in processing time is accounted for by the five predictors. The four word-information measures together account for 41.96%.

[c] The strong correlation between syntactic surprisal and syntactic entropy reduction seems to be an artifact of the artificial nature of the language. On a corpus of newspaper texts, the correlation between the two information measures was found[13] to be only around .25.

There is still 56.99% of variance in processing time unaccounted for. In order to investigate to what extent this may be due to imperfections in the SRN's output, the dynamical process of Eq. 8 was run once more but this time on the correct situation vectors $\mu(\mathrm{sit}(w_{1\ldots i}))$ rather than the network outputs $\mathbf{a}_{\mathrm{out}}(i)$. As shown in Table 4, this increases the fraction of total variance accounted for to as much as 83.68%, suggesting that a large part of the variance that is not accounted for in Table 3 is indeed caused by noise in the network's output.

Table 4. Regression analysis results when using correct situation vectors instead of network outputs. Factors are in order of inclusion in the regression model.

Predictor	Coefficient	R^2
s_{sem}	0.50	.758
ΔH_{sem}	4.66	.075
Position	0.30	.003
ΔH_{syn}	0.34	.001
s_{syn}	0.30	.000

5. Discussion

The sentence-comprehension model implements the mental-simulation theory of language understanding, in the sense that the comprehension process results in a non-linguistic, analogical representation of the situation described by the sentence. This representation is constructed incrementally: It is continuously updated as the sentence's words come in one by one. The incorporation of the incoming word into the current representation continues until the representation's rate of change drops below a threshold level. Consequently, more time is required if the word signals a greater change in the situation being described. As such, it may not come as a surprise that processing times correlate with formal measures of semantic information content: Words that convey much information are precisely those that have a large impact on the described situation.

Nevertheless, the current findings have implications for theories about the role of formal notions of information in cognition. Both surprisal and entropy reduction have been claimed to quantify amount of information and as such were assumed to predict word-reading times. However, this does not explain *why* information content would be related to processing time. The fact that surprisal and entropy reduction have independent effects on human word-reading times[13,14] would seem to indicate that there are two distinct cognitive processes or representations involved, one explaining the effect of surprisal and the other that of entropy reduction. The current finding, however, suggest an alternative: The only relevant comprehension processes is the revision of a single mental representation. Surprisal and entropy reduction merely form two complementary quantifications of the extent of this revision. The fact that they are formal measures of information does not give them any special cognitive status. Although language understanding can be viewed as information processing, information-theoretic concepts do not necessarily correspond to psycholinguistic reality.

In addition to semantic information, other factors were found to affect word-processing time. First, the word's position in the sentence had a positive effect. This contradicts experimental findings which have shown that reading speeds up over the course of a sentence.[8] Hence, the model cannot explain this effect of word position. Possibly, the construction of a mental simulation is not responsible for faster reading at later words.

Second, it is of some psycholinguistic interest that syntactic surprisal and syntactic entropy reduction accounted for some of the variance in word-processing time. Again, the relation was positive: words that convey more syntactic information are processed more slowly. Somehow, the network has become sensitive to the statistical patterns in the microlanguage and is thereby able to account (at least to some extent) for the effects of syntactic information on cognitive processing effort, which have been found in human reading-time data. When word-processing times were obtained by using the correct situation vectors rather than the network outputs, the effect of syntactic information all but disappeared: As can be seen in Table 4, the total R^2 associated with syntactic information is now only 0.001. The effect of syntactic surprisal is no longer significant ($p > .9$). This suggests that the relation between syntactic information and processing time is not just a fluke of the particular microlanguage and microworld, but originates from the network that learns to map the sentences to the corresponding situation vectors. Interestingly, this task does not require any syntactic knowledge, that is, knowledge of sentences probabilities is not needed for learning the form-meaning mapping. The finding that syntactic knowledge nevertheless affects word-processing times therefore suggests that learning the meaning of language can result in the acquisition of syntax as a side effect.

6. Conclusion

It was shown how cognitively relevant definitions of syntactic information can be extended to semantics. All it takes is a shift in focus from the statistics of the language (syntactic patterns) to the statistics of the world (semantic patterns). A connectionist sentence-comprehension model, rooted in ideas from Embodied Linguistics, predicted that words that convey more information take longer to process, irrespective of the information source (linguistic knowledge or world knowledge) and information measure (surprisal or entropy reduction). This constitutes a first step towards a more computationally developed theory of Embodied Linguistics, in which the incorporation of world knowledge into the definition of information provides a more sound formal basis to the notion of mental simulation.

Acknowledgements

This research was supported by grant 277-70-006 of the Netherlands Organization for Scientific Research (NWO). I am grateful to Michael Klein and Eddy Davelaar for commenting on an earlier version of this paper.

References

1. J. Hale, A probabilistic Early parser as a psycholinguistic model, in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, (Pittsburgh, PA: Association for Computational Linguistics, 2001) pp. 159–166.
2. J. Hale, *Journal of Psycholinguistic Research* **32**, 101 (2003).
3. R. Levy, *Cognition* **106**, 1126 (2008).
4. R. Zwaan, Experiential traces and mental simulations in language comprehension, in *Symbols, embodiment, and meaning: debates on meaning and cognition*, eds. M. D. Vega, A. M. Glenberg and A. C. Graesser (Oxford, UK: Oxford University Press, 2008) pp. 165–180.
5. D. J. Chwilla, H. H. J. Kolk and C. T. W. M. Vissers, *Brain Research* **1183**, 109 (2007).
6. J. Hale, *Cognitive Science* **30**, 643 (2006).
7. M. F. Boston, J. Hale, U. Patil, R. Kliegl and S. Vasishth, *Journal of Eye Movement Research* **2**, 1 (2008).
8. V. Demberg and F. Keller, *Cognition* **109**, 193 (2008).
9. B. Roark, A. Bachrach, C. Cardenas and C. Pallier, Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing, in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, (Association for Computational Linguistics, Singapore, 2009) pp. 324–333.
10. H. Brouwer, H. Fitz and J. Hoeks, Modeling the noun phrase versus sentence

coordination ambiguity in Dutch: Evidence from surprisal theory, in *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, (Association for Computational Linguistics, Uppsala, Sweden, 2010) pp. 72–80.

11. S. L. Frank, Surprisal-based comparison between a symbolic and a connectionist model of sentence processing, in *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, eds. N. A. Taatgen and H. van Rijn (Austin, TX: Cognitive Science Society, 2009) pp. 1139–1144.
12. N. J. Smith and R. Levy, Optimal processing times in reading: a formal model and empirical investigation, in *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, eds. B. C. Love, K. McRae and V. M. Sloutsky (Austin, TX: Cognitive Science Society, 2008) pp. 595–600.
13. S. L. Frank, Uncertainty reduction as a measure of cognitive processing effort, in *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, (Association for Computational Linguistics, Uppsala, Sweden, 2010) pp. 81–89.
14. S. Wu, A. Bachrach, C. Cardenas and W. Schuler, Complexity metrics in an incremental right-corner parser, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, (Association for Computational Linguistics, Uppsala, Sweden, 2010) pp. 1189–1198.
15. B. Kaup, R. H. Yaxley, C. J. Madden, R. A. Zwaan and J. Lüdtke, *Quarterly Journal of Experimental Psychology* **60**, 976 (2007).
16. S. L. Frank, W. F. M. Haselager and I. Van Rooij, *Cognition* **110**, 358 (2009).
17. S. L. Frank, M. Koppen, L. G. M. Noordman and W. Vonk, *Cognitive Science* **27**, 875 (2003).
18. L. W. Barsalou, *Behavioral and Brain Sciences* **22**, 577 (1999).
19. R. A. Stanfield and R. A. Zwaan, *Psychological Science* **12**, 153 (2001).
20. J. L. Elman, *Cognitive Science* **14**, 179 (1990).