Information **2011**, *2*, 672-696; doi:10.3390/info2040672

OPEN ACCESS

information

ISSN 2078-2489

www.mdpi.com/journal/information

Article

Sentence Comprehension as Mental Simulation: An Information-Theoretic Perspective

Stefan L. Frank * and Gabriella Vigliocco

Department of Cognitive, Perceptual and Brain Sciences, University College London, 26 Bedford Way, London WC1H 0AP, UK; E-Mail: g.vigliocco@ucl.ac.uk (G.V.)

* Author to whom correspondence should be addressed; E-Mail: s.frank@ucl.ac.uk; Tel.: +44-2076795589.

Received: 2 July 2011; in revised form: 25 October 2011 / Accepted: 17 November 2011 / Published: 23 November 2011

Abstract: It has been argued that the mental representation resulting from sentence comprehension is not (just) an abstract symbolic structure but a "mental simulation" of the state-of-affairs described by the sentence. We present a particular formalization of this theory and show how it gives rise to quantifications 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 dynamic connectionist model of sentence comprehension as mental simulation. A quantitatively similar relation between information content and reading time is known to be present in human reading-time data.

Keywords: sentence comprehension; mental simulation; word information; connectionist modeling; word-reading time; semantic and syntactic bootstrapping

1. Introduction

1.1. Representing Meaning

In most cognitive models of language comprehension (e.g., [1–3]), the meaning of a sentence is represented as a structural combination of concepts, forming a predicate-argument structure. For example, the semantic representation of *Heidi plays chess* may be something like play (heidi, chess) or (action:play, agent:heidi, patient:chess). Even many connectionist (or "neural network") models,

which are often claimed to be non-symbolic and therefore of a fundamentally different nature, make use of predicate-argument structures when they represent the meaning of a proposition by concatenating (*i.e.*, structurally combining) activation patterns that form arbitrary (*i.e.*, symbolic [4]) representations of the concepts involved (e.g., [5–8]).

Predicate-argument structures are compositional representations: Their interpretation is a function of the symbolic elements of which they are composed and the manner in which these elements are structurally combined. The idea that human cognition has such a combinatorial syntax and semantics is known as the Language of Thought hypothesis [9] because language itself is (at least to a considerable extent) compositional too. Note that predicate-argument structures are indeed language-like: The words of *Heidi plays chess* correspond one-to-one to the concepts of play(heidi, chess), and the semantic structure follows the sentence's hierarchical syntactic structure.

An alternative approach to the representation of meaning is to follow the structure of the *world* rather than the compositional structure of language, as in Johnson-Laird's [10] Mental Model theory and, more recently, Barsalou's [11] Perceptual Symbol System hypothesis. Such representations do not consist of amodal symbols but are modal (*i.e.*, they have sensorimotor qualities) and analogical, which is to say that the relation between a representation's form and its meaning is not arbitrary. According to this view, to understand a sentence is to *mentally simulate* the state-of-affairs it describes [12,13]. The representation of *Heidi plays chess*, for example, would be similar to the result of actually observing Heidi playing chess. In the current paper, we will present a formal representational scheme that captures the analogical (but not the modal) character of perceptual symbols.

An increasing amount of empirical evidence supports the idea that comprehension is mental simulation. In one experiment [14], for example, participants who were told that *the pencil is in the cup* automatically and unconsciously formed a mental representation of a pencil in a vertical position, as evidenced by faster response time to a picture of a vertically oriented pencil than to the same picture rotated by 90°. The outcome was reversed when the participants heard that *the pencil is in the drawer*. Such results show that to mentally represent a sentence's meaning is also to represent what the sentence implies. If mental representations were merely symbolic predicate-argument structures, a special inference process would be needed to get from in(pencil, cup) to orientation(pencil, vertical). However, it would remain unexplained why this particular inference is drawn, considering that the experimental participants had no reason to pay attention to the pencil's orientation. To explain the experimental results, therefore, it must be assumed that (nearly) all possible inferences are drawn. Moreover, the visual stimulus would need to be encoded as orientation(pencil, vertical), rather than simply as pencil, even though the pencil's orientation is one of its many irrelevant properties. Analogical representations, in contrast, can do without such prolific and explicit inference because the representation of a pencil in a cup *is also* a representation of a pencil in vertical orientation.

There are only a few computational models that implement language comprehension as mental simulation, possibly because it is very difficult to formalize and define operations over the required representations (the highly complex DETE model [15] is a case in point). Embodied Construction Grammar [16] takes mental simulation to operate over symbolic rather than analogical representations and, consequently, requires an explicit process to generate inferences from a semantic specification of the input sentence. In contrast, the formalization we present here does use analogical representations. It

was originally developed for the Distributed Situation Space (DSS) model [17] of story comprehension, which was able to account for a range of experimental findings regarding inference and recall in discourse comprehension [17] and the resolution of ambiguous pronouns [18]. In addition, the connectionist sentence-comprehension model by Frank, Haselager, & Van Rooij [19] generates DSS representations and has demonstrated how internalizing the structures of the language and the world results in systematic semantic processing, even without compositionality or structure-sensitive operations. In the following sections, we will show how the Frank *et al.* model naturally gives rise to different, complementary formalizations of the amount of information words convey, and use these word-information measures to account for word-processing times.

1.2. Language Comprehension as Information Processing

Mental-simulation theory may seem at odds with more formal, information-theoretic approaches to language processing, which have so far only been concerned with the statistics of linguistic structures rather than the states of the world referred to. Nevertheless, we will connect the mental-simulation and information-theoretic views by demonstrating how internalized knowledge of the structure of the world can lead to formal measures of the amount of information conveyed by the words of a sentence.

Information-theoretic measures of word-information with respect to the structure of the *language* (rather than the world) are already well known. Such measures can be computed by probabilistic models that capture the statistical patterns in a large corpus of sentences. Assuming that humans have learned the same (or at least very similar) statistics, these information-theoretic measures should predict cognitive processing effort: 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 [20–23].

Linguistic patterns are a part of the overall statistics of the world. It is therefore only natural to extend the information-theoretic view of sentence processing to incorporate non-linguistic statistical patterns. The idea here is that humans have internalized much of these statistics, just as they have for the statistics of language, and use this knowledge for mental simulation in language comprehension. The sentence-comprehension process should therefore be affected by non-linguistic information in the same way as by linguistic information. Indeed, world knowledge affects language processing at an early stage during comprehension [24,25]. Processing difficulty occurs when construction of the mental simulation is more challenging. 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 [26], the N400 component on the sentence-final word was found to be larger in (1b) than in (1a), indicating relative comprehension difficulty in (1b). This could not simply be attributed to differences in our experience with linguistic forms: According to an analysis of co-occurrences of words in text corpora, the two sentences are equally novel. A possible explanation of the effect is that mentally simulating the described action is more difficult in (1b) than in (1a), because it makes less 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 (non-linguistic) world [27] that is responsible for the N400 effect in this experiment.

Similar effects occur in reading times and cannot simply be attributed to semantic or associative priming [28–31]. These findings raise the interesting possibility that the amount of cognitive effort required for processing a word (as observed in its reading time) can be quantified by information measures that are based on the sentence's meaning in relation to our knowledge of the world, rather than the sentence's form (or structure) in relation to our knowledge of the language. As we shall see, in our particular implementation of mental-simulation theory, linguistic information measures extend naturally to meaning.

1.3. Overview

The main purpose of the current paper is to present world-knowledge based formalizations of the amount of information conveyed by each word in a sentence. These information measures follow from a sentence-comprehension model that is rooted in the mental-simulation view of language understanding. We extend this model with a dynamical process that allows it to generate predictions of word-processing time, which should be predicted by the word-information measures.

As we have argued above, knowledge of (and experience with) the real world is fundamental to language comprehension. Our modeling exercise, however, is necessarily restricted in size and scope. As such, the model only deals with a small "microworld", states of which are described by sentences of a "microlanguage". Consequently, the model's processing-time predictions cannot be quantitatively compared to human experimental data, nor are we able to analyze particular psycholinguistic phenomena or linguistic structures. Instead, we will only investigate qualitative correspondences between model results and human data.

In the remainder of the paper, we first present Frank *et al.*'s [19] sentence-comprehension model and explain how it is used here to simulate word-reading times. Next, Section 3 formalizes the notion of the amount of information conveyed by words, both with respect to linguistic knowledge and world knowledge. If the sentence-comprehension process is indeed sensitive to information quantities, the information-theoretic measures should be predictive of word-processing times. Indeed, as shown in Section 4, the model takes longer to process words that convey more information, be it based on linguistic or world knowledge. Section 5 discusses these findings, in particular their relevance to human language processing and acquisition, and Section 6 concludes.

2. The Sentence-Comprehension Model

As an implementation of mental-simulation theory, the Frank *et al.* model [19] treats word-by-word comprehension of a sentence as the incremental construction of a representation of the described state-of-affairs. As explained below (and in more detail in Appendix A.2), these representations capture the probabilities of states of the world, making the framework ideally suited for incorporating measures of word-information with respect to world knowledge, and for studying their effect on sentence processing. A simple extension of the model makes it possible to obtain word-processing times, which will be compared to the different word-information measures.

2.1. The Microworld and Microlanguage

According to the mental-simulation view of language comprehension, understanding a sentence relies fundamentally on real-world knowledge and experience. To make the amount of knowledge manageable for the model, the world is restricted to a small microworld, which we will only present briefly here since details can be found elsewhere [19].

The microworld has just three inhabitants (sophia, heidi, and charlie) and four locations (bedroom, bath, 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.

It is important to keep in mind that these predicate-argument structures are only used here for communicative purposes. As will become clear in Section 2.2, there are no symbolic or compositional representations within the model. Instead, microworld situations are represented analogically.

Some situations are more likely to occur than others. To name a few microworld tendencies, heidi is prone to win at hide&seek, sophia usually loses at chess, and charlie is most often in a different place than the girls. There also exist hard constraints on possible situations. Although a few of these are somewhat arbitrary (e.g., the ball is only played with in outside locations), most hard constraints pose reasonable restrictions on the microworld's "metaphysical" possibilities. For example, each of the three protagonists is always in exactly one place and someone has to play some game in order to win or lose.

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)

Microworld situations are described by microlanguage sentences. Again, full details are published elsewhere [19] so they will not be presented here. The language has a vocabulary of 40 words, including 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. Situations that violate microworld constraints can be described by grammatical microlanguage sentences (e.g., *ball is played with in bedroom*) but there are no sentences for situations that are logically or metaphysically impossible, such as win(charlie) $\wedge\neg$win(charlie) or place(sophia,street) $\wedge$ place(sophia,playground).

2.2. Representing Microworld Situations

As mentioned in the Introduction, a number of psycholinguistic experiments have shown that the mental representation resulting from sentence comprehension mirrors the state-of-affairs described by the sentence, which may include things that are not literally stated in (but are inferable from) the sentence. In the context of our microworld, for example, to represent Sophia playing with the ball is also to represent her being outside. The comprehension model's representations, being analogical rather than symbolic, have exactly this property. Thus, such inferences in the model can be considered as emergent properties: They come about as a consequence of the need to represent situations in an analogical manner.

Microworld situations are represented by vectors in a 100-dimensional "situation space". Each situation in the microworld corresponds to one vector in this space. Conversely, each situation-space vector (or "situation vector" for short) corresponds to some state of the microworld, albeit rarely one that can be neatly described as a boolean combination of atomic situations.

The vectors representing atomic situations are automatically extracted from a "large enough" sample of 25,000 examples of microworld states, randomly generated by an algorithm that takes all the microworld's constraints and correlations into account. Each such example shows which atomic situations are the case (and which are not) at one moment in microworld time, and the patterns across these examples embody the probabilistic constraints on co-occurrences of microworld situations. The extraction of situation vectors from the examples efficiently re-encodes the co-occurrence patterns into situation space (see Appendix A.1 for details). As a result, each atomic situation p is represented by a vector $\mu(p)$ whose average element value is a (highly accurate) estimate of the probability that p is the case at any particular moment in the microworld. As we discuss in Section 3, this encoding will turn out to be instrumental when defining measures of semantic information.

Crucially, the vector representations, like the atomic events themselves, can be combined using boolean operators, yielding yet more situation vectors that capture the situations' probabilities. For example, $\mu(p \wedge q)$ represents the conjunction "p and q" by virtue of the fact that its average element value accurately estimates the probability that p and q occur simultaneously in the microworld.

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 estimated directly from $\mu(p)$ and $\mu(q)$. Such estimates are called *belief values* and denoted by the symbol τ. That is, the belief value $\tau(p|q)$ is an estimate of $P(p|q)$, the probability that p is the case given that q. For example, if charlie plays soccer he must be playing with the ball, and indeed $\tau(\mathsf{play}(\mathsf{charlie},\mathsf{ball})|\mathsf{play}(\mathsf{charlie},\mathsf{soccer})) > 0.99$. This shows that a situation-space representation of some fact also forms a representation of anything the fact implies, which is precisely what makes this representational scheme a model of analogical mental representations. Also, note how this scheme is fully in line with recent theories concerning the probabilistic nature of human cognition [32].

2.3. The Comprehension Process

Comprehension of a microlanguage sentence is modeled as the process of mapping the sentence onto the vector representing the described microworld situation. This mapping is learned and performed by a Simple Recurrent Network (SRN) [33]; a standard connectionist architecture for incremental sentence processing that has been used to account for a variety of phenomena in human language acquisition and comprehension ([5–7,34–37]). The network has a three-layer architecture (see Figure B1 in Appendix B), with one input unit for each word in the microlanguage, one output unit for each dimension of situation space, and a number of hidden units that is fairly arbitrarily set to 75. When a sentence that describes microworld situation p has been processed, the network's output ideally equals p's vector representation $\mu(p)$, which therefore forms the target output when the network is trained on that sentence. Before the sentence is completed, the output should reflect what is described so far. That is, if p, q, r, ... are all the microworld situations consistent with the sentence-so-far, the network's output vector ideally equals $\mu(p \vee q \vee r \vee \ldots)$, expressing the disjunction of all situations that are not yet excluded by the incoming sentence. During network training, however, the target output is $\mu(p)$ from the very beginning of the sentence.

One drawback of the standard SRN is that it has no notion of processing a word over time: Processing each input always takes one sweep of activation through the network, irrespective of the input's identity or context. Consequently, the network cannot directly simulate word-processing times. To extract reading-time predictions from the network, processing time must be made to depend on the "amount of work" involved in processing each particular word in its particular context. This is accomplished by turning the network's output update function into a dynamical process, which means that processing a word involves changing the output activation pattern over continuous time rather than instantaneously (for mathematical details, see Appendix B). This process halts when the rate of change, which decreases over time, falls below a particular threshold level. At that point, the resulting output is (nearly) the same as it would be in the original, non-dynamical model.

Only the SRN's output is updated dynamically, so the time required to process the word depends only on the word's effect on the network's output pattern. In other words, the only process that is assumed to take any time is the update of the mental simulation. This is, of course, a simplification. In reality, word-reading times also depend on linguistic/syntactic processing and, in fact, the model does predict syntactic effects on processing times (see the results in Section 4). Moreover, it has been argued [38] that the mental representation of a statement's meaning can also contain non-simulated, symbolic or linguistic elements.

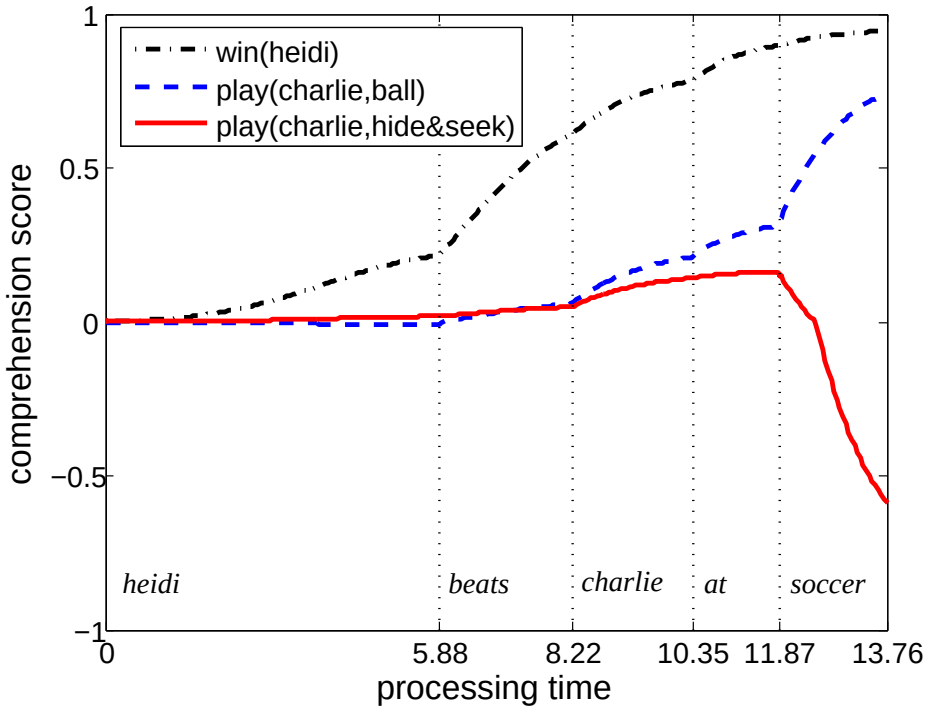
2.4. Sentence Comprehension Example

We illustrate the model's sentence-comprehension process by presenting simulation results on a single example sentence. Directly observing the network's output activation (*i.e.*, the situation vector) as it develops over processing time is not particularly informative. Instead, we interpret the output's meaning by computing *comprehension scores* for particular situations of interest. As defined in [19], the comprehension score for a situation p is a value between -1 and $+1$ that indicates the extent to which a (possibly incomplete) sentence asserts that p is the case. The comprehension score is $+1$ if p has a belief

value τ of 1 in the situation represented by the current output vector. Likewise, the score is -1 if p has a belief value of 0. If p's belief value equals its prior probability (e.g., at the beginning of a sentence, before any information about the state of the world is provided), its comprehension score equals 0.

Figure 1 shows comprehension scores during processing of the sentence *heidi beats charlie at soccer* for three atomic situations of interest. The first of these, win(heidi), is literally asserted by the sentence. The second, play(charlie, ball), is not literally asserted but, by applying microworld knowledge, can be inferred from what the sentence claims. The third, play(charlie, hide&seek), is inconsistent with the situation described by the sentence because Charlie cannot be playing both soccer and hide-and-seek.

Figure 1. Comprehension scores of three situations during processing of *heidi beats charlie at soccer*. The labels on the horizontal axis indicate the time points (in arbitrary units) at which a new word enters the model.



As is clear from Figure 1, the model indeed creates a situation vector that accurately represents the state of the world as described by the sentence. Moreover, it does so incrementally: The situation vector is continuously updated to reflect what has been understood from the sentence so far. For example, the second word, *beats*, strongly increases the belief value of win(heidi). It keeps increasing over the course of the sentence because new words can enter the model before the previous words have been fully processed. The next word is *charlie*, telling us that he must be playing some game. As a result, the belief values of play(charlie, ball) and play(charlie, hide&seek) increase slightly, that is, the model can use its knowledge of the microworld to anticipate the upcoming message. When the last word arrives, we learn that Charlie was playing soccer so play(charlie, ball) is considered increasingly likely whereas the amount of belief in play(charlie, hide&seek) diminishes.

3. Quantifying Word Information

As discussed in the Introduction, the language and the world form two different sources of statistical patterns with respect to which information measures can be defined. We will call these *syntactic* information and *semantic* information, respectively. It is important to keep in mind that we use these terms in a very specific sense, which may be slightly different from their use in (psycho)linguistics. By *syntax* we refer to the probabilities of word strings, where grammatical sentences are precisely those strings with strictly positive probability. That is, we do not mean to evoke any particular grammar from which sentence probabilities (or even just grammaticality judgments) arise. Our use of *semantics* is restricted to the relation between a sentence and the state-of-affairs in the world to which it refers. That is, we do not attempt to relate our model to the representations and analyses that are common in the field of formal semantics. Also, we only make a strict *analytical* distinction between syntactic and semantic information here, in order to clarify the current modeling work. We do not wish to claim that these two information source are also *cognitively* distinct. As we said before, language is part of our world so there is no a priori reason to assume that linguistic knowledge and world knowledge are acquired using fundamentally different learning processes.

An orthogonal distinction can be made between two different definitions of the amount of information conveyed by each word of a sentence, which have both been proposed in the computational psycholinguistics literature: *surprisal* [20,23] and *entropy reduction* [21,22,39]. The first is a formal measure of the extent to which a word came unexpected, whereas the latter quantifies the extent to which a word reduces uncertainty about the upcoming material.

This section explains how these four information measures (syntactic surprisal, semantic surprisal, syntactic entropy reduction, and semantic entropy reduction) can be defined in general and how their specific values follow from the microworld and microlanguage discussed in the previous section.

3.1. Surprisal

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

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

How can the unexpectedness of a word's occurrence be formally quantified? Suppose the cognitive language-processing system has perfect knowledge of the language's syntax, meaning that it has access to the true probability of every sentence. Also, it does perfect incremental sentence processing, meaning that after processing each word it excludes all sentences that are inconsistent with the input string so far. Now let $P(w_{1\ldots n})$ denote the probability of the n-word sentence $w_{1\ldots n}$ (which is short for w_1, w_2, …, w_n) in the language. After the first i words (*i.e.*, $w_{1\ldots i}$) of the current input sentence have been processed, the sentence probabilities will have changed. In particular, any sentence that does not begin with $w_{1\ldots i}$

will now have zero probability. The probability that the sentence will turn out to be $w_{1...n}$ given the sentence-so-far, equals:

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

Having assumed perfect knowledge of the sentence probabilities, $P(w_{1...i}, w_{i+1...n})$ is known. The probability of the sentence-so-far, $P(w_{1...i})$ is simply the total probability of all sentences that begin with those words:

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

This shows that the hypothetical language-processing system is able to compute the probability distribution over all sentences given the sentence so far. Also, it can generate expectations about what the following word is going to be: The probability that this is some particular word w_{i+1}, equals:

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

A word's *syntactic surprisal* is an information-theoretic measure that expresses the extent to which the word's occurrence was not expected, considering knowledge of sentence probabilities. Syntactic surprisal follows directly from the word's probability given its sentence context:

$$s_{\text{syn}}(w_{i+1}) = -\log P(w_{i+1}|w_{1...i}) \quad (1)$$

The negative logarithm makes sure that surprisal values can range from zero (when no other word could have occurred, *i.e.*, $P(w_{i+1}|w_{1...i}) = 1$) to infinite (when the word's occurrence was considered impossible, *i.e.*, $P(w_{i+1}|w_{1...i}) = 0$) [40]. Syntactic surprisal is easy to compute in our microlanguage, because the complete set of sentences and their probabilities are known. In more realistic cases, a probabilistic language model needs to be trained on a text corpus after which it can be used to estimate surprisal values over novel sentences.

It has been suggested that syntactic surprisal is indicative of cognitive processing effort and should therefore be predictive of word-reading time [20,23]. Indeed, reading times have repeatedly been shown to correlate positively with syntactic surprisal, as estimated by a wide variety of probabilistic models of language [36,41–47].

We called this measure "syntactic" because it is defined with respect to the probabilities of sentences, and does not depend on what these sentences mean. The task of our hypothetical sentence-processing system was merely to identify the incoming sentence. In contrast, if the system has knowledge of (the probabilities of) the many possible states of the world and knows which sentence refers to which (set of) situation(s), a notion of *semantic surprisal* arises. Semantic surprisal follows from the statistical patterns of the world rather than the language. To give a simple example, the situation described in (3a) is more likely than the one in (3b) according to our knowledge of academia:

(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 leads to the assertion of a situation that is unlikely to occur, given what was already learned from the sentence so far.

Assume that each sentence $w_{1...n}$ refers to a situation in the world. This situation, denoted $\text{sit}(w_{1...n})$ occurs with a probability of $P(\text{sit}(w_{1...n}))$. The first i words of a sentence describe the situation $\text{sit}(w_{1...i})$, which is the disjunction of all situations described by sentences that start with those words. For instance, $\text{sit}(\textit{Heidi plays}) = \text{sit}(\textit{Heidi plays chess}) \vee \text{sit}(\textit{Heidi plays with ball}) \vee \text{sit}(\textit{Heidi plays outside}) \vee \ldots$. The sentence's next word, w_{i+1}, changes the situation from $\text{sit}(w_{1...i})$ to $\text{sit}(w_{1...i+1})$. 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 Equation (1):

$$s_{\text{sem}}(w_{i+1}) = -\log P(\text{sit}(w_{1...i+1})|\text{sit}(w_{1...i})). \tag{2}$$

Note that the definition of semantic surprisal does not use the probabilities of the word strings $w_{1...i}$ and $w_{1...i+1}$. Only the probabilities of the described situations are relevant. In the context of the sentence-comprehension model presented in Section 2, the situations are states of the microworld and the conditional probability in Equation (2) is estimated by the belief value $\tau(\text{sit}(w_{1...i+1})|\text{sit}(w_{1...i}))$, which follows directly from the vector representations of the two situations involved.

This quantification of a word's semantic surprisal is a special case of Bar-Hillel and Carnap's [48] definition of the amount of semantic information conveyed by one statement over and above what is conveyed by some other statement. Unlike that notion of semantic information, Equation (2) puts strict constraints on the two statements, in that the one must be expressed by the first i words of a sentence and the other by the first $i + 1$ words of the same sentence. This turns Bar-Hillel and Carnap's more general definition into one that is tailored to incremental sentence comprehension.

3.2. Entropy Reduction

The second measure of the amount of information conveyed by a word formalizes the idea that each word usually (but not necessarily) decreases the amount of uncertainty about what is being communicated. 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 as informative about what the next words will be, so more uncertainty remains. Hence, the word "cats" conveys more syntactic information in (2a) than in (2b).

Again, this notion can be quantified by hypothesizing an ideal sentence-processing system. After processing the words $w_{1...i}$, how certain can such a system be about the identity of the current sentence? If only one sentence is compatible with the input so far, the probability of that sentence will be 1 while all other probabilities are 0. In contrast, uncertainty is maximal when all sentences have the same probability (except for those that are inconsistent with $w_{1...i}$). The information-theoretic notion of *entropy* quantifies uncertainty by looking at the shape of the probability distribution. If there is one outcome with probability 1, entropy is 0; If probability is distributed evenly over possible outcomes,

entropy is maximal. The syntactic entropy over possible sentences given the input $w_{1...i}$ so far, is formally expressed by:

$$H_{\text{syn}}(i) = - \sum_{w_{1...i}, w_{i+1...n}} P(w_{1...i}, w_{i+1...n}|w_{1...i}) \log P(w_{1...i}, w_{i+1...n}|w_{1...i}) \quad (3)$$

The *syntactic entropy reduction* due to processing the next word w_{i+1} equals:

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

and has been hypothesized to be a cognitively relevant measure of the amount of information conveyed by the word [21,22,39]. Although the entropy-reduction hypothesis has not been tested as extensively as the effect of syntactic surprisal, the two syntactic word-information measures have been shown to be independent predictors of word-reading times [49,50].

Entropy reduction, like surprisal, can also be defined with respect to knowledge of the world. To get an idea of what the resulting *semantic entropy reduction*, ΔH_{sem}, would mean, compare the following sentence fragments:

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

Brilliant papers get accepted whereas the fate of a mediocre paper is unsure. Therefore, the uncertainty about the situation being communicated is higher after processing fragment (4a) than (4b). In either sentence, the final word could turn out to be "accepted" or "rejected". At that point, there is (arguably) no more uncertainty, so semantic entropy is zero. Consequently, the final word (be it "accepted" or "rejected") reduces uncertainty more strongly (*i.e.*, conveys more semantic information) in (4a) than in (4b).

One major difficulty when formalizing semantic entropy reduction is that the computation of entropy requires access to a complete probability distribution. For syntactic entropy, this is the probability distribution over sentences, which is known in our sentence-comprehension model and can to some extent be estimated in realistic cases ([45,49]). For semantic entropy, however, we need the probability distribution over all possible, fully specified states of the world. Even in our tiny microworld, obtaining this distribution would require probability estimates for $2^{44} \approx 10^{13}$ boolean combinations of atomic events, which is clearly infeasible.

This problem is solved by replacing the set of fully specified situations by a much smaller set of mutually exclusive, partially unspecified situations, and computing entropy over their probabilities. The resulting approximate entropy-reduction measures should at least form reasonable estimates of the true (but unknowable) values. In our current modeling framework, choosing the set of partially unspecified situations is surprisingly straightforward: We can simply use the 100 dimensions of situation space themselves, because these can be expected to form a reasonable encoding of everything there is to know about the microworld. Each dimension corresponds to some microworld situation (which is unlikely to be verbalizable or fully specified) and has a belief value given the situation described by the sentence so far. It is easy to prove (see Appendix C) that the resulting collection of 100 belief values indeed forms a probability distribution.

3.3. Word Information Example

Table 2 presents the four different amounts of information conveyed by each word of the example sentence from Section 2.4. Note that the amounts of syntactic and semantic information differ strongly. In particular, the function word *at* conveys only syntactic, not semantic, information. This is because another word could have occurred at that position (so *at* conveys syntactic information) whereas it was already clear that Heidi must be winning *at* something from the fact that she beats someone. Therefore, the occurrence of *at* tells us nothing new about the state of the microworld. More surprisingly, perhaps, is the finding that the sentence-initial word *heidi* conveys no semantic information either. The reason is that the mere mention of *heidi* does not constrain the possible microworld situations. If the sentence had started with *soccer* it would have constrained the possible situations to those in which soccer is being played, thereby conveying semantic information.

Table 2. The syntactic and semantic information conveyed by each word in *heidi beats charlie at soccer*.

Measure	***heidi***	***beats***	***charlie***	***at***	***soccer***
s_{syn}	1.99	0.73	1.60	1.25	1.60
ΔH_{syn}	1.99	0.67	1.61	1.31	1.60
s_{sem}	0	2.05	0.90	0	1.58
ΔH_{sem}	0	1.78	1.00	0	1.52

4. Results

Everything is now in place to investigate the relation between the 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. The pairwise correlations between these information measures are shown in Table 3, which also includes the position of the word in the sentences as a covariate. The correlation between the two syntactic measures is very strong [51] and, consequently, one measure cannot explain much variance in word-processing time over and above the other. Therefore, we leave syntactic entropy reduction out of the analysis.

Table 3. Matrix of correlation coefficients between factors.

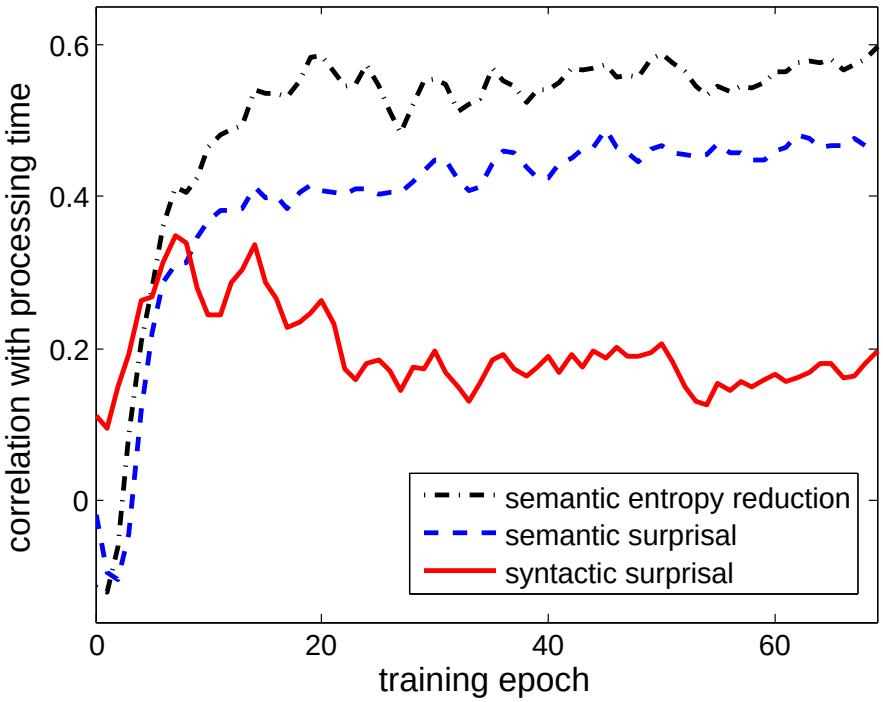
	s_{sem}	ΔH_{syn}	ΔH_{sem}	**Position**
s_{syn}	0.09	0.95	0.24	−0.03
s_{sem}		0.09	0.28	0.24
ΔH_{syn}			0.24	−0.07
ΔH_{sem}				0.08

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 [52].

Figure 2 shows the correlation coefficients between word-processing times and the remaining three information measures, over the course of network training. Note that the network appears to become sensitive to syntactic patterns early in learning, before semantic information has much of an effect. Later, the two semantic information-measures dominate, and the effect of syntactic information diminishes. It is clear that each information measure correlates positively with reading times: Words that convey more information, be it syntactic or semantic, take longer to process. For syntactic information, this is in line with much experimental data [36,41–47]. For semantic information, this result is what we would expect considering that the network is engaged in semantic (more than syntactic) processing. Also, several psycholinguistic experiments [28–31] have shown that reading times are indeed longer on words that are less appropriate considering the state-of-affairs being described and that can therefore be considered semantically more informative.

Figure 2. Coefficients of correlation between word-processing times and information measures, over network training time. Each training epoch corresponds to the presentation of 30,000 sentence tokens. For presentation purposes, the correlations are smoothed by averaging over a 3-epoch window.



The results in Figure 2 do not show whether each information measure accounts for variance over and above all the others, and also the effect of word position is not taken into account. To investigate more thoroughly whether different word-information measures truly have independent predictive value, we performed a stepwise regression analysis, also including the position of the word in the sentence as a predictor. This is done at three moments during network training: At the point where syntactic information has maximal effect (epoch 7), at the point where it seems to have reached a stable level

(epoch 25), and after complete training (epoch 69). Table 4 presents the results, showing each predictor's coefficient and the fraction of variance it explains (R^2) over and above what is accounted for by the predictors already in the regression model. Predictors were added to the regression one by one, in order of decreasing R^2. Each addition resulted in a significant ($p < 0.001$) increase in regression model fit.

Table 4. Regression analysis results: coefficients and fractions of unique variance explained.

Predictor	Epoch 7		Epoch 25		Epoch 69	
	Coefficient	R^2	**Coefficient**	R^2	**Coefficient**	R^2
s_{syn}	0.309	0.093	0.071	0.002	0.086	0.002
s_{sem}	0.006	0.030	0.014	0.068	0.022	0.103
ΔH_{sem}	0.240	0.196	0.528	0.328	0.699	0.355
Position	0.041	0.009	0.015	0.001	0.061	0.005

The regression results confirm what was already suggested by Figure 2: Each of the word-information measures contributes positively to word-processing time, so a word takes longer to process if it conveys more syntactic or semantic information. Early in training, there is still a considerable effect of syntactic surprisal (accounting for 9.3% of variance in processing time) but when training is completed its contribution is tiny (0.2%), albeit still statistically significant. At that point, the two semantic information measures combined explain as much as 45.8% of variance in processing time.

5. Discussion

Our sentence-comprehension model implements the mental-simulation theory of language understanding, in the sense that the outcome of the comprehension process is a non-linguistic, analogical representation of the situation described by the sentence. This representation is constructed incrementally: As the sentence's words come in one at a time, the situation vector is continuously updated to provide the best match to all the available evidence. Such immediate integration of information sources to converge on the best fitting interpretation is known as *constraint satisfaction* in the psycholinguistics literature [53]. Hence, the model can be considered to provide a constraint-satisfaction account of mental simulation. In addition to immediate integration, the model also displays the ability to use its world knowledge to anticipate the upcoming message: Processing *heidi beats charlie* resulted in increased belief values of play(charlie, ball) and play(charlie, hide&seek) (see Figure 1). Such anticipatory inference also occurs in human language comprehension, as has been shown in studies using the visual-world paradigm [54,55].

5.1. The Effect of Information Content on Processing Time

When a word enters the model, the current representation of the described situation changes continuously until the rate of change drops below a certain level, at which point the word is considered sufficiently processed. Consequently, more time is required if the word signals a greater change in the situation being described, as has been observed in human readers [28–31]. Considering that words

that convey more semantic information are precisely those that have a larger impact on the described situation, it may not come as a surprise that the model's word-processing times correlate with formal measures of semantic information content. Nevertheless, it is relevant that the two measures of semantic information have independent effects on word-processing time. Human reading-time data has shown independent effects of *syntactic* surprisal and entropy reduction [49,50] so we would expect the same for the semantic measures, if knowledge about the statistics of the world is acquired and used in the same manner as knowledge about the statistics of language.

Interestingly, syntactic surprisal was also found to affect word-processing time, in particular in the earlier stages of network training. When the network has become sensitive to the statistical patterns in the microlanguage, words that convey more syntactic information are processed more slowly, as has been observed in human reading-time data [36,41–47]. This effect diminishes as training proceeds, which comes as no surprise since the sentences' probabilities are irrelevant to the network's task of mapping form to meaning: Ideally, the network's output to any input string $w_{1...i}$ always equals the corresponding situation vector $\mu(\text{sit}(w_{1...i}))$, irrespective of the probability of $w_{1...i}$.

This raises the question why an effect of syntactic surprisal appears at all. The model will acquire knowledge about the statistics of linguistic patterns (and, thereby, possibly display an effect of syntactic surprisal) only to the extent that it is helpful to its semantic task. Be reminded that the network is trained to generate the situation vector corresponding to the *complete* sentence from the very first word onwards. Of course, after having seen just the incomplete sentence $w_{1...i}$ it cannot yet come up with $\mu(\text{sit}(w_{1...n}))$, the vector representing the situation described by the complete input $w_{1...n}$ (unless no other situation is consistent with $w_{1...i}$). The best it can do, therefore, is to generate $\mu(\text{sit}(w_{1...i}))$. However, if the network is able to anticipate upcoming word(s), it can make a reasonable guess at the correct output $\mu(\text{sit}(w_{1...n}))$. Hence, learning which words are (syntactically) likely to occur after $w_{1...i}$ is useful because it speeds up processing of those words, even though the task of mapping sentences onto situation vectors does not require syntactic knowledge. Presumably, this is why there is an initial, positive effect of syntactic surprisal on word-processing time. As learning proceeds, and the network becomes more attuned to the rarer linguistic patterns, the output to $w_{1...i}$ comes to resemble $\mu(\text{sit}(w_{1...i}))$ more closely, increasing the effect of semantic information at the expense of syntactic surprisal.

5.2. Relation to Human Language Acquisition

The initial decrease in correlation between semantic information and processing time (see Figure 2) may seem to suggest that (some) syntax is learned first and semantics follows, just like infants can develop sensitivity to syntactic patterns before they show any semantic understanding [56]. However, this would be a misinterpretation of the network's behavior: No part of the model is dedicated to learning purely linguistic patterns. Rather, all it learns is to incrementally map sentences onto their semantic representations, so all the syntactic knowledge it picks up must be by virtue of learning what the sentences mean. Indeed, the network's acquisition of semantics is not delayed: Comprehension performance improves from the very first training epoch, as revealed by an immediate decrease in the difference between the network's output and the correct situation vectors. However, for reasons unknown, word-processing times are not immediately affected by semantic information.

Considering that the network is designed to learn only semantics, yet becomes sensitive to syntax, it must engage in some form of *semantic bootstrapping*: the acquisition of syntactic knowledge from associations between form and meaning [57,58]. Although semantic bootstrapping is most commonly viewed as a mechanism by which infants can learn the syntactic categories of words, the model shows that it may also be applicable at the sentence level: Learning the meaning of sentences results in sensitivity to the language's statistics, as apparent from the effect of syntactic surprisal on word-processing time.

In addition, the model's early use of linguistic patterns, before semantics seems to play a role, is reminiscent of *syntactic bootstrapping* [59,60]: the acquisition of the meaning of language from syntactic structure (which we have simplified to linguistic patterns of occurrence). There is ample experimental support for the occurrence of syntactic bootstrapping in first language acquisition [61,62]. If we would wish to argue that our model displays or simulates something like syntactic bootstrapping, it needs not only be shown that the early use of syntactic information improves or speeds up sentence comprehension, but also that it facilitates the *acquisition* of semantics. Although the current results do not tell us whether that is indeed the case, it may be possible (at least in principle) to redesign the microlanguage such that syntactic knowledge is no longer relevant, for example by replacing each sentence by a unique symbol, making the language non-compositional. If the network indeed uses linguistic patterns to boost the acquisition of semantics, learning should slow down when syntactic information cannot be used.

In reality, of course, words do convey syntactic information. A properly tuned language processing system can use that to its advantage, anticipating upcoming input and thereby speeding up the comprehension process. It is noteworthy in this respect that our microlanguage and microworld showed a positive correlation between the syntactic and semantic information measures (see Table 3). This, we believe, is not just coincidental: Since language is often used to describe a part of reality, the statistics of language and of the world are unlikely to be fully independent. We therefore expect that more realistic sets of linguistic and world knowledge would also show a correlation between syntactic and semantic information.

If language and the world indeed constitute interdependent (rather than independent) sources of knowledge, learning from their combination should be more successful than learning from the two independently. At lexical level, it has indeed been shown that richer semantic representations arise when the words' patterns of occurrence in linguistic data and their perceptual/experiential features are treated as an integrated data set, rather than as two independent sources [63,64]. If the same is true at the sentence level, the structure and meaning of language should be learned more fully, or more quickly, when linguistic data and world data are integrated. That is, a language-acquisition system should perform better if (unlike our model) it learns about linguistic and non-linguistic structures simultaneously, using a single mechanism.

6. Conclusion

We have presented one of the first computational models that treats sentence comprehension as mental simulation. It was shown how the model naturally leads to cognitively relevant notions of semantic information, based on well-known definitions of syntactic information. All it takes is a shift in focus from the statistics of the language (syntactic patterns) to the statistics of the world (semantic patterns).

The model predicted that it takes longer to process words that convey more information, irrespective of the information source (linguistic knowledge or world knowledge) and information measure (surprisal or entropy reduction).

It remains to be seen whether all these information measures indeed have independent behavioral effects in human readers. Such an investigation requires probabilistic formalizations of semantics and world knowledge that go far beyond those of the toy language and world presented here. The current work constitutes a first step towards a more computationally developed theory of perceptual symbol systems: By incorporating world knowledge into the definition of word information, we have provided a formal basis to the notion of mental simulation.

Acknowledgments

The research presented here was funded by the European Union 7th Framework Programme (FP7/2007-2013) under grant number 253803. This is an extended and improved version of a paper that appeared as: Frank, S.L. (2011). Sentence comprehension as mental simulation: an information-theoretic analysis and a connectionist model. In: E.J. Davelaar (Ed.), *Connectionist models of neurocognition and emergent behavior: Proceedings of the 12th Neural Computation and Psychology Workshop* (pp. 190–207). Singapore: World Scientific.

References and Notes

1. Budiu, R.; Anderson, J.R. Interpretation-based processing: A unified theory of semantic sentence comprehension. *Cogn. Sci.* **2004**, *28*, 1–44.
2. Kintsch, W. The role of knowledge in discourse comprehension: A construction-integration model. *Psychol. Rev.* **1988**, *95*, 163–182.
3. Padó, U.; Crocker, M.W.; Keller, F. A probabilistic model of semantic plausibility in sentence processing. *Cogn. Sci.* **2009**, *33*, 794–838.
4. Peirce, C.S. Logic as Semiotics: The Theory of Signs. In *Semiotics: An Introductory Anthology*; Innis, R.E., Ed.; Indiana University Press: Bloomington, IN, USA, 1903/1985; pp. 4–23.
5. Desai, R. Bootstrapping in miniature language acquisition. *Cogn. Syst. Res.* **2002**, *3*, 15–23.
6. Desai, R. A model of frame and verb compliance in language acquisition. *Neurocomputing* **2007**, *70*, 2273–2287.
7. Mayberry, M.R.; Crocker, M.W.; Knoeferle, P. Learning to attend: A connectionist model of situated language comprehension. *Cogn. Sci.* **2009**, *33*, 449–496.
8. Morris, W.C.; Cottrell, G.W.; Elman, J. A Connectionist Simulation of the Empirical Acquisition Of Grammatical Relations. In *Hybrid Neural Systems*; Wermter, S., Sun, R., Eds.; Springer: London, UK, 2000; pp. 177–193.
9. Fodor, J.A. *The Language of Thought*; Harvard University Press: Cambridge, MA, USA, 1975.
10. Johnson-Laird, P.N. *Mental Models*; Cambridge University Press: Cambridge, UK, 1983.
11. Barsalou, L.W. Perceptual symbol systems. *Behav. Brain Sci.* **1999**, *22*, 577–660.
12. Glenberg, A.M.; Robertson, D.A. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *J. Mem. Lang.* **2000**, *43*, 379–401.

13. Zwaan, R. Experiential Traces and Mental Simulations in Language Comprehension. In *Symbols, Embodiment, and Meaning: Debates on Meaning and Cognition*; Vega, M.D., Glenberg, A.M., Graesser, A.C., Eds.; Oxford University Press: Oxford, UK, 2008; pp. 165–180.
14. Stanfield, R.A.; Zwaan, R.A. The effect of implied orientation derived from verbal context on picture recognition. *Psychol. Sci.* **2001**, *12*, 153–156.
15. Nenov, V.I.; Dyer, M.G. Perceptually grounded language learning: Part 2–DETE: A neural/procedural model. *Connect. Sci.* **1994**, *6*, 3–128.
16. Bergen, B.K.; Chang, N.C. Embodied Construction Grammar in Simulation-Based Language Understanding. In *Construction Grammars: Cognitive Grounding and Theoretical Extensions*; Östman, J., Fried, M., Eds.; John Benjamins: Amsterdam, The Netherlands, 2005; pp. 147–190.
17. Frank, S.L.; Koppen, M.; Noordman, L.G.M.; Vonk, W. Modeling knowledge-based inferences in story comprehension. *Cogn. Sci.* **2003**, *27*, 875–910.
18. Frank, S.L.; Koppen, M.; Noordman, L.G.M.; Vonk, W. Coherence-driven resolution of referential ambiguity: A computational model. *Mem. Cogn.* **2007**, *35*, 1307–1322.
19. Frank, S.L.; Haselager, W.F.M.; van Rooij, I. Connectionist semantic systematicity. *Cognition* **2009**, *110*, 358–379.
20. Hale, J. A Probabilistic Early Parser as a Psycholinguistic Model. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA, USA, 2–7 June 2001; Volume 2, pp. 159–166.
21. Hale, J. The information conveyed by words. *J. Psycholinguist. Res.* **2003**, *32*, 101–123.
22. Hale, J.T. What a rational parser would do. *Cogn. Sci.* **2011**, *35*, 399–443.
23. Levy, R. Expectation-based syntactic comprehension. *Cognition* **2008**, *106*, 1126–1177.
24. Hagoort, P.; Hald, L.; Bastiaansen, M.; Petersson, K.M. Integration of word meaning and world knowledge in language comprehension. *Science* **2004**, *304*, 438–441.
25. van Berkum, J.J.A.; Brown, C.M.; Zwitserlood, P.; Kooijman, V.; Hagoort, P. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *J. Exp. Psychol. Learni. Mem. Cogn.* **2005**, *31*, 443–467.
26. Chwilla, D.J.; Kolk, H.H.J.; Vissers, C.T.W.M. Immediate integration of novel meanings: N400 support for an embodied view of language comprehension. *Brain Res.* **2007**, *1183*, 109–123.
27. From hereon, we leave implicit that by "world" we mean the non-linguistic part of the world.
28. Bicknell, K.; Elman, J.L.; Hare, M.; McRae, K.; Kutas, M. Effects of event knowledge in processing verbal arguments. *J. Mem. Lang.* **2010**, *63*, 489–505.
29. Camblin, C.C.; Gordon, P.C.; Swaab, T.Y. The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *J. Mem. Lang.* **2007**, *56*, 103–128.
30. Matsuki, K.; Chow, T.; Hare, M.; Elman, J.L.; Scheepers, C.; McRae, K. Event-based plausibility immediately influences on-line language comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* **2011**, *37*, 913–934.
31. Traxler, M.J.; Foss, D.J.; Seely, R.E.; Kaup, B.; Morris, R.K. Priming in sentence processing: Intralexical spreading activation, schemas, and situation models. *J. Psycholinguist. Res.* **2000**, *29*, 581–595.

32. Oaksford, M.; Chater, N. *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*; Oxford University Press: Oxford, UK, 2007.
33. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211.
34. Christiansen, M.H.; Chater, N. Toward a connectionist model of recursion in human linguistic performance. *Cogn. Sci.* **1999**, *23*, 157–205.
35. Christiansen, M.H.; MacDonald, M.C. A usage-based approach to recursion in sentence processing. *Lang. Learn.* **2009**, *59*, 129–164.
36. Frank, S.L. Surprisal-Based Comparison Between a Symbolic and a Connectionist Model of Sentence Processings. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Amsterdam, The Netherlands, 29 July–1 August 2009; Taatgen, N.A., van Rijn, H., Eds.; pp. 1139–1144.
37. Rohde, D.L.T. A Connectionist Model of Sentence Comprehension and Production. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2002.
38. Hemforth, B.; Konieczny, L. Language Processing: Construction of Mental Models or More? In *Mental Models and the Mind*; Held, C., Knauff, M., Vosgerau, G., Eds.; Elsevier: Amsterdam, The Netherlands, 2006; pp. 189–204.
39. Hale, J. Uncertainty about the rest of the sentence. *Cogn. Sci.* **2006**, *30*, 643–672.
40. If the system has perfect knowledge of the language, a word that it considers impossible will (by definition) never occur, so surprisal cannot reach infinity. In practice, when there is no perfect knowledge, infinite surprisal is prevented by smoothing the word probabilities so that they are always greater than zero.
41. Boston, M.F.; Hale, J.; Patil, U.; Kliegl, R.; Vasishth, S. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *J. Eye Mov. Res.* **2008**, *2*, 1–12.
42. Boston, M.F.; Hale, J.; Vasishth, S.; Kliegl, R. Parallel processing and sentence comprehension difficulty. *Lang. Cogn. Process.* **2011**, *26*, 301–349.
43. Demberg, V.; Keller, F. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* **2008**, *109*, 193–210.
44. Frank, S.L.; Bod, R. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychol. Sci.* **2011**, *22*, 829–834.
45. Roark, B.; Bachrach, A.; Cardenas, C.; Pallier, C. 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*, Singapore, 6–7 August 2009; pp. 324–333.
46. Smith, N.J.; Levy, R. Optimal Processing Times in Reading: A Formal Model and Empirical Investigation. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, Washington, DC, USA, July 2008; Love, B.C., McRae, K., Sloutsky, V.M., Eds.; pp. 595–600.
47. Wu, S.; Bachrach, A.; Cardenas, C.; Schuler, W. Complexity Metrics in an Incremental Right-Corner Parser. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11–16 July 2010; pp. 1189–1198.
48. Bar-Hillel, Y.; Carnap, R. Semantic information. *Br. J. Philos. Sci.* **1953**, *4*, 147–157.

49. Frank, S.L. Uncertainty Reduction as a Measure of Cognitive Processing Effort. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, Uppsala, Sweden, July 2010; pp. 81–89.
50. Frank, S.L. Uncertainty reduction as a measure of cognitive effort in sentence comprehension. *Manuscript in preparation*.
51. This is probably an artifact of the artificial nature of the language or of the fact that the true sentence probabilities are known rather than estimated by a model. On a corpus of newspaper texts, the correlation between the two syntactic information measures was found to be only around .25 [49] and it was even lower on a random selection of sentences from novels [50].
52. The correlations coefficients of Table 3 are based on this reduced data set.
53. MacDonald, M.C.; Seidenberg, M.S. Constraint Satisfaction Accounts of Lexical and Sentence Comprehension. In *Handbook of Psycholinguistics*, 2nd ed.; Traxler, M., Gernsbacher, M., Eds.; Academic Press: London, UK, 2006; pp. 581–611.
54. Kamide, Y.; Altmann, G.T.M.; Haywood, S.L. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *J. Mem. Lang.* **2003**, *49*, 133–156.
55. Knoeferle, P.; Crocker, M.W. The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye-tracking. *Cogn. Sci.* **2006**, *30*, 481–529.
56. Marcus, G.F.; Vijayan, S.; Rao, S.B.; Vishton, P.M. Rule learning by seven-month-old infants. *Science* **1999**, *283*, 77–80.
57. Pinker, S. *Language Learnability and Language Development*; Harvard University Press: Cambridge, MA, USA, 1984.
58. Pinker, S. The Bootstrapping Problem in Language Acquisition. In *Mechanisms of Language Acquisition*; MacWhinney, B., Ed.; Erlbaum: Hillsdale, NJ, USA, 1987; pp. 399–441.
59. Gleitman, L. The structural sources of verb meanings. *Lang. Acquis.* **1990**, *1*, 3–55.
60. Gleitman, L.R.; Cassidy, K.; Nappa, R.; Papafragou, A.; Trueswell, J.C. Hard words. *Lang. Learn. Dev.* **2005**, *1*, 23–64.
61. Naigles, L.R.; Swensen, L.D. Syntactic Supports for Word Learning. In *Blackwell Handbook of Language Development*; Hoff, E., Shatz, M., Eds.; Blackwell Publishing Ltd.: Hoboken, NJ, USA, 2008; pp. 212–231.
62. Fisher, C.; Gertner, Y.; Scott, R.M.; Yuan, S. Syntactic bootstrapping. *WIREs Cogn. Sci.* **2010**, *1*, 143–149.
63. Andrews, M.; Vigliocco, G.; Vinson, D. Integrating experiential and distributional data to learn semantic representations. *Psychol. Rev.* **2009**, *116*, 463–498.
64. Steyvers, M. Combining feature norms and text data with topic models. *Acta Psychol.* **2010**, *3*, 234–342.

A. Situation Space

A.1. Constructing Situation Space

Vector representations of the 44 atomic microworld situations were obtained by training a so-called *competitive layer* on 25,000 examples of situations occurring in the microworld. Each of these examples takes the form of a 44-element binary vector, containing 0s and 1s to indicate which basic events are (not) the case at some moment in microworld time. These were sampled by a non-deterministic algorithm that goes through the 44 atomic situations in a random order and sets each value to 0 or 1, according to the atomic situation's probability given the truth values of the atomic situations so far.

The competitive layer, consisting of 100 units, takes the microworld examples as input and self-organizes to represent the fact that p is the case in the microword as an activation pattern $\mu(p) = (\mu_1(p), \ldots, \mu_{100}(p)) \in [0,1]^{100}$ over its units. Details of the competitive layer learning algorithm can be found in [19]. Ten competitive layers were trained, each on a different sample of 25,000 microworld examples, resulting in ten different situation spaces.

A.2. Estimating Probabilities of Situations

Each situation vector encodes the probability that the represented situation occurs in the microworld. As it turns out, the probability of a situation closely approximates its vector's average element value:

$$P(p) \approx \frac{1}{100} \sum_j \mu_j(p) \equiv \tau(p) \tag{4}$$

Moreover, the average value of the elementwise product of two vectors approximates the probability that the two represented situations co-occur:

$$P(p \wedge q) \approx \frac{1}{100} \sum_j \mu_j(p)\mu_j(q) \equiv \tau(p \wedge q) \tag{5}$$

It follows from the relation between vectors and probabilities, as expressed in the two equations above, that any complex microworld situation can be represented as a vector in situation space by simple operations on the atomic vectors. First, since $P(\neg p) = 1 - P(p)$, it follows from Equation (4) that $\mu(\neg p) = 1 - \mu(p)$. Second, from Equation (5) it follows that $\mu_j(p \wedge q) = \mu_j(p)\mu_j(q)$. Third, the vector representing the disjunction $p \vee q$ can be constructed by making use of the fact that $p \vee q \equiv \neg(\neg p \wedge \neg q)$. In short, any complex situation can be represented by boolean combinations of the 44 atomic situation vectors, and the probability of the situation always approximates the vector's average element value. Consequently, the probability of any (atomic or complex) situation p, given that q is the case, can be observed directly from $\mu(p)$ and $\mu(q)$:

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

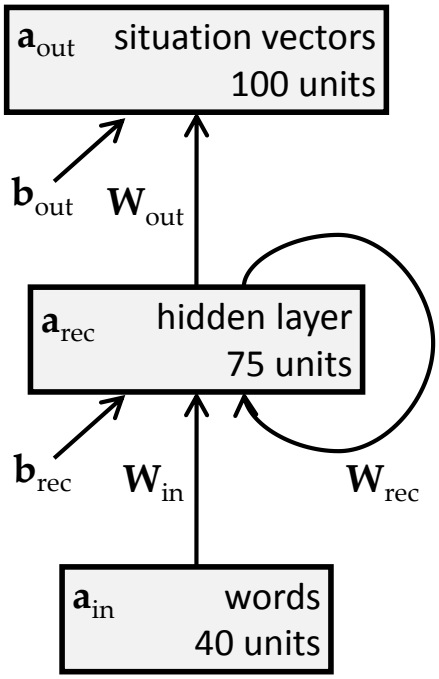
In each of the ten situation spaces, estimated probabilities correlated very strongly with the "true" probabilities in the total set of 10 $\times$ 25,000 examples, indicating that situation vectors indeed capture microworld probabilities very accurately: The correlation between $\tau(p|q)$ and $P(p|q)$, over all 44 atomic events p and all 88 (negations of) atomic events q, was always above 0.998.

B. The Sentence-Comprehension Network

B.1. Network Processing

Figure B1 shows the SRN architecture of the sentence-comprehension network. Each input unit corresponds to a word of the microlanguage, and the output activations $\mathbf{a}_{\text{out}}$ are interpreted as situation vectors.

Figure B1. Architecture of the sentence-comprehension network.



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

$$\begin{aligned}\mathbf{a}_{\text{rec}}(i+1) &= f(\mathbf{W}_{\text{rec}}\mathbf{a}_{\text{rec}}(i) + \mathbf{W}_{\text{in}}\mathbf{a}_{\text{in}}(i+1) + \mathbf{b}_{\text{rec}}) \\ \mathbf{a}_{\text{out}}(i+1) &= f(\mathbf{W}_{\text{out}}\mathbf{a}_{\text{rec}}(i+1) + \mathbf{b}_{\text{out}})\end{aligned} \tag{6}$$

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: Rather than taking Equation (6), we make the smallest possible change that turns it in into a simple differential equation:

$$\frac{\mathrm{d}\mathbf{a}_{\text{out}}}{\mathrm{d}t} = f(\mathbf{W}_{\text{out}}\mathbf{a}_{\text{rec}}(i+1) + \mathbf{b}_{\text{out}}) - \mathbf{a}_{\text{out}} \tag{7}$$

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. (This is because the unit vector conveys no information about the state of the microworld: $\tau(p|\mathbf{1}) = \tau(p)$ for any p. Therefore, before the first word of the sentence is processed, $\mathbf{a}_{\text{out}}$ encodes that all microworld situations occur with their prior probabilities.) Equation (7) expresses that the output vector $\mathbf{a}_{\text{out}}$ moves over processing time from $\mathbf{a}_{\text{out}}(i)$ 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 Equation (6). This process converges when $\mathbf{a}_{\text{out}}$ no longer changes, that is, when $\mathrm{d}\mathbf{a}_{\text{out}}/\mathrm{d}t = 0$, which 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 (Equation (6)). 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:

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

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 Equation (8) is the simulated reading time on word w_{i+1}.

B.2. Network Training

Training examples were randomly sampled from all 13,556 microlanguage sentences, with shorter sentences having a larger sampling probability. The probabilities of sentences are chosen independently from the probabilities of the described events in the world. However, longer (*i.e.*, less probable) sentences tend to specify more details about the state of the world, and more specified situations have lower probability. As a result, there is an $r = 0.34$ correlation between the probabilities of sentences and the probabilities of the corresponding situations.

Each sampled sentence was presented to the network, one word at a time. After each word, the network's actual output is compared to the target output, consisting of the vector representation of the situation described by the complete sentence. Connection weights are updated according to the standard backpropagation algorithm, with a learning rate of 0.02.

After every 30,000 training sentences, the network's outputs to all complete sentences are compared to the corresponding targets, and the mean squared error between the two is computed. The network is considered fully trained as soon as this error falls below 2% of the pre-training error (*i.e.*, the error in the output of a network with random weights).

Ten networks were trained, differing in their initial random weight settings and the training sentences' sampling probabilities. Also, the networks were paired one-to-one with the 10 situation spaces, so each had a different set of target outputs. Qualitatively similar results were obtained across the different networks. We present only the results of the one network that was closest to the average (over the 10 networks) in the total variance in processing time accounted for by the four predictors (three information measures and word position, as in Table 4).

C. Semantic Entropy Reduction

Ideally, semantic entropy would follow from the probability distribution over all states of the world. That is, after processing the sentence-so-far $w_{1\ldots i}$, the entropy would be

$$H_{\text{sem}}(i) = -\sum_{S \in \mathcal{S}} P(S|\text{sit}(w_{1\ldots i})) \log P(S|\text{sit}(w_{1\ldots i}))$$

where $\mathcal{S}$ denotes the set of all possible, fully specified situations. However, even in our simple microworld, it is virtually impossible to list all these situations and their probabilities.

An obvious solution is to replace $\mathcal{S}$ by a much smaller set $\mathcal{S}'$ of situations that are not fully specified. This needs to be done with care though, because situations that are not fully specified may not be mutually exclusive. For example, the situation in which play(heidi, chess) and the one in which place(charlie, street) are different but do not exclude each other since it is possible that play(heidi, chess) $\wedge$ place(charlie, street) occurs. As a result, the collection of probabilities over incompletely specified situations may sum to more than one, and therefore not form a probability distribution. Conversely, if $\mathcal{S}'$ does not cover all possible situations, the total probability will be less than one, again violating a constraint on probability distributions.

So the question becomes: How to choose the set $\mathcal{S}'$ such that its elements are mutually exclusive and their disjunction covers the complete microworld? A trivial answer is to choose one atomic situation, say p, and have just two situations in $\mathcal{S}'$: One in which p is the case, and one in which it is not. This is obviously not a acceptable solution.

Rather than coming up with a reasonable $\mathcal{S}'$ ourselves, we leave this to the competitive layer that develops the situation vectors (see Appendix A.1). Let $\mu(p_x)$ denote the situation vector consisting of 0s except that its xth element is 1, that is:

$$\mu_j(p_x) = \begin{cases} 0 & \text{if } x \neq j \\ 1 & \text{if } x = j \end{cases}$$

The vector represents some state p_x of the microworld, but it is unlikely that this situation can be expressed precisely as a boolean combination of atomic events. No matter what p_x refers to, however, its probability given the sentence so far can be estimated by its belief value:

$$\tau(p_x|\text{sit}(w_{1\ldots i})) = \frac{\sum_j \mu_j(p_x)\mu_j(\text{sit}(w_{1\ldots i}))}{\sum_j \mu_j(\text{sit}(w_{1\ldots i}))} = \frac{\mu_x(\text{sit}(w_{1\ldots i}))}{\sum_j \mu_j(\text{sit}(w_{1\ldots i}))}$$

The total belief value of the p_x's for each of the 100 situation-space dimensions x equals

$$\sum_{p_x} \tau(p_x|\text{sit}(w_{1\ldots i})) = \sum_x \frac{\mu_x(\text{sit}(w_{1\ldots i}))}{\sum_j \mu_j(\text{sit}(w_{1\ldots i}))} = 1$$

so the $\tau(p_x|\text{sit}(w_{1\ldots i}))$ form a proper probability distribution. Also, it is clear that the different p_x are mutually exclusive since $\tau(p_x \wedge p_y) = 0$ for $x \neq y$.

The competitive layer attempts to find situation-space dimensions that are maximally informative about the probabilistic structure of the microworld. Consequently, the set $\mathcal{S}' = \{p_x\}$ (for $x = 1, \ldots, 100$) can be expected to form a reasonable alternative to $\mathcal{S}$. Hence, we take as semantic entropy measure:

$$H_{\text{sem}}(i) = -\sum_{p_x \in \mathcal{S}'} \tau(p_x|\text{sit}(w_{1\ldots i})) \log \tau(p_x|\text{sit}(w_{1\ldots i}))$$

The semantic entropy reduction due to the next word, w_{i+1}, is simply

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