Information-theoretical complexity metrics

John Hale
Department of Linguistics
Cornell University

February 2, 2016
Abstract

Information-theoretical complexity metrics are auxiliary hypotheses that link theories of parsing and grammar to potentially observable measurements such as reading times and neural signals. This review article considers two such metrics, Surprisal and Entropy Reduction, which are respectively built upon the two most natural notions of ‘information value’ for an observed event [Blachman, 1968]. The review sketches their conceptual background, and touches on their relationship to other theories in cognitive science. It highlights their status as lenses through which the consequences of grammatical representations can be seen, and it suggests mechanistic interpretations for both of them.

1 One of many applications

Since the 1990s, information theory in the sense of Shannon [1948] has enjoyed a revival right across the language sciences. As a mathematical theory, it applies quite generally to questions about language structure, acquisition and processing. Malouf [forthcoming] provides an extensive bibliography. This review focuses more narrowly on complexity metrics (see section 2 below) that relate theories of syntactic parsing to empirical data such as reading times or blood oxygen levels in the brain.

Delimiting the focus in this way, I nonetheless encourage readers to approach information theory on its own terms; John R. Pierce’s [1980] provides a nontechnical book-length introduction. Kornai [2007, §7.2] is a shorter, more mathematical treatment embedded in a broader discussion of linguistic complexity. Readers interested in phonological applications should consult Goldsmith [2000] or Hume and Mailhot [2013]. For morphology, see Milin et al. [2009]. Information theory also figures prominently in models of sentence misperception [Levy, 2008b, Gibson et al., 2013], language production [Keller, 2004, Jaeger, 2010] and language learning [Chater et al., 2015]. This modeling literature is sometimes polemical, arguing for a “rational analysis” of human psychology founded on a Bayesian interpretation of probability. In what follows, I set aside this polemic since information theory is compatible with any philosophy of probability.

2 Complexity metrics

Generally speaking, a complexity metric is something that quantifies how difficult it is to perceive a linguistic expression. This characterization includes any rule that relates theorized parsing mechanisms to word-by-word measures of comprehension difficulty. Kaplan’s [1972] “number of transitions made or attempted” (page 89) is a classic example. But there are many others: the number of unresolved case-roles at a given point [Stabler, 1994, Morrill, 2000], the number of phrase structure nodes in memory [Yngve, 1960, Frazier, 1985], or the latency of a memory retrieval operation motivated by a syntactic dependency [Lewis and Vasishth, 2005]. The length of this catalog testifies to the important role that complexity metrics have always played in relating theorized mechanisms to
observable data. One commonality among the metrics just cited is the fact that they make per-word, rather than per-sentence predictions. It is this “incremental” feature that is important for online comprehension studies. By contrast, the ordered list of construction types given in Caplan et al. [1985, 123] or the ranking of larger and smaller domains suggested in Hawkins [2010] both offer their predictions at the level of the sentence, rather than the word. They are complexity metrics but not incremental ones.

3 Deriving predictions about potential observations

Surprisal and Entropy Reduction are incremental complexity metrics that predict how difficult each word should be as it is perceived in time. They are information-theoretical insofar as they view sentences as random events. From this technical perspective, words are symbols that appear with a certain probability. The revelation of each new word-symbol as being the particular symbol that it is, constitutes a sub-event with a quantifiable information value. Both metrics suppose that greater information value should relate to greater processing difficulty. However they differ in the precise formulation of ‘information value’ that they apply. Both can be thought of as summaries of the transition between a word and its successor, but they are mathematically different and can derive contrasting empirical consequences e.g. on relative clauses (see section 5.4). In typical psycholinguistic modeling practice, one computes the value of the metric at each word-position in a stimulus sentence. These predictions are then compared with observed measures of comprehension effort, such as reading times, scalp potentials or blood oxygen levels in particular brain areas. If the per-word information values match up well with the observed measures then we say that the observed data support the conjunction of the complexity metric and whatever defined the probabilities in the first place.

Of course the key issue is what defines the probabilities. This boils down to the question of language model, to which we now turn.

4 Language Models

Information-theoretical complexity metrics like Surprisal and Entropy Reduction are defined in terms of probability distributions having to do with the transition from one word to the next. The specification of these distributions is called a language model. The word ‘language’ in this name alludes to the classic conceptualization of a language as a set of strings of words [e.g. Chomsky, 1956]. In this narrow sense, a language model is simply an assignment of probability to each string in this set. The question is, how to define these probabilities?

4.1 The obvious way is vulnerable to a famous critique

Perhaps the most straightforward way to assign probabilities to strings is to probabilistically generate each word, one after the other. Under this arrangement, the probability of a successor word is defined in terms of the previous words that have already been generated. The general view is that of a table keeping track of (a) the last \( n - 1 \) words, (b) the \( n^{\text{th}} \) word and (c) its conditional probability, \( P(w_i \mid w_{i-1} w_{i-2} \cdots w_{i-(n-1)}) \). Rows of this table, namely combinations of possible successor words sharing the same left-context, serve to define a discrete distribution. One can then ask how surprising is a specific word \( w_i \), how uncertain is the distribution as a whole, et cetera.

This view is appealingly straightforward, but as a scientific proposal it leaves much to be desired. In fact, it is exactly the Markov model of language that Chomsky [1956] criticized. The core of
Chomsky’s critique is that, in real human language, words influence each other at arbitrary distances. In cases of grammatical agreement, coordination or relative clause formation, for instance, the presence or absence of an upcoming word depends upon words that may be quite far back in the stream. If the dependency is not fulfilled, the probability should be essentially zero. For dependencies wider than $n$, it will be impossible for an $n$th-order Markov model to assign realistic probabilities. The problem is not a matter of degree. Rather, it is that the Markov property itself that fails to be fulfilled by natural language.

### 4.2 The conditional probability of a grammatical derivation

To overcome this problem, speech recognition pioneers like Jelinek and Lafferty [1991] turned to probabilistic grammars. In a probabilistic grammar, conditional probabilities are associated with rewriting rules, so that derivation is now a branching process as shown in Figure 1.

Figure 1: Two derivations that share the same initial substring (bold words). Downward arrows emphasize that derivation is a stochastic process. The sum of the probabilities of all derivations that share an initial substring is the probability that the grammar assigns to that initial substring.

Viewing derivation as a branching process in this way means that derivation trees are values of a random variable, $X$. Conditioning on a particular initial substring, such as **john loves**, isolates a particular subset of these derivations. This is analogous to selecting rows in a conditional probability table, except that the subset may itself be infinite. A probabilistic grammar defines a distribution on this subset just as it does for the entire language. It is this distribution, or rather, before/after pairs of distributions on either side of the latest word, that information-theoretical metrics summarize. The next section applies this perspective to Surprisal, before moving on to Entropy Reduction in section 6.

---

3For a gentle introduction to probabilistic grammars see Figure 1 in Chater and Manning [2006] or chapter 14 of Jurafsky and Martin [2008].
5 Surprisal

5.1 Definition

The term “surprisal” dates back to Tribus [1961] who used it to talk about the logarithm of the reciprocal of a probability. It is one way to characterize the information value of an observed event (section 6 presents another). For a generic random variable \( Y \), the surprisal of an outcome \( Y = y \) is:

\[
\log_2 \left( \frac{1}{P(y)} \right)
\]

Note that the argument to the logarithm in this expression is the reciprocal of a probability. Since probabilities are between 0 and 1, the smallest value of this reciprocal corresponds to the largest probability i.e. \( \frac{1}{1} \). As probabilities get closer to zero, their reciprocal gets larger and larger. Figure 2 shows this pattern using a dashed line. Taking the logarithm of these reciprocals pulls the curve down a bit but retains the idea of assigning higher values to lower probabilities.\(^4\) In a nutshell, surprisal is higher for low-probability events.\(^5\)

To apply this formal definition, one needs to fix upon a particular event: the appearance of a word as the next symbol in a string. This implies the existence of two strings, both anchored at the beginning of a sentence with one is exactly one symbol longer than the other. These are “initial substrings” of the sort indicated in bold in Figure 1. Both of them may be associated with conditional probability distributions using a grammar. As suggested in that Figure, the support of these distributions is precisely the set of derivations that derive them. The total probability mass that they assign is known as the “prefix probability,” and it is the ratio of these prefix probabilities that gives the transition probability of the next symbol. This ratio defines \( P(y) \) in the surprisal complexity metric. This is written out below in equation 2 where these prefix probabilities are

\[\text{Figure 2: Reciprocal and surprisal in the interval (0, 1)}\]

\(^4\)The choice of a logarithmic measure is defended in the Introduction to Shannon’s Mathematical Theory of Communication [1948]. Specifying that the base of the logarithm is 2 is a calibration which means that the surprisals are counted in bits, just like computer memory. Through linear regression one can fit these information theoretical predictions in bits to particular dependent variables, e.g. reading times in milliseconds.

\(^5\)This logarithmic difficulty rule is analogous to the Hick-Hyman law, which characterizes the time a person takes to decide between multiple choices. See Pierce [1980, 230].
cartooned as sums (\(\sum\)) either before the successor word or after it.

\[
\log_2 \left( \frac{1}{P(y)} \right) \quad \text{let } y \text{ be the ratio of prefix probabilities}
\]

\[
= \log_2 \left( \frac{1}{\sum_{\text{after}}/\sum_{\text{before}}} \right)
\]

\[
= \log_2 \left( \frac{\sum_{\text{before}}}{\sum_{\text{after}}} \right)
\]

\[
= -\log_2 \left( \frac{\sum_{\text{after}}}{\sum_{\text{before}}} \right)
\] (2)

To illustrate this application of the generic surprisal idea in equation 1 to grammatical derivations, as in equation 2, Figure 3 shows a pair of hypothetical probability distributions of the sort that might be generated by a probabilistic grammar. Each bar corresponds to a derivation, such as those in Figure 1. The height of the bars represents each derivation’s probability. The histogram on the left depicts the distribution at the shorter initial substring. With the appearance of the next word, the set of available derivations contracts. Some derivations are incompatible with the new word, as shown in the histogram on the right which corresponds to the longer substring. Successive transitions zero-out derivations in this way until, in an unambiguous sentence, presumably only one is left. Each time this happens, the probability assigned to the missing bars is lost.\(^6\) The total amount lost corresponds logarithmically to the predicted comprehension difficulty.

Figure 3: Transitioning from word to word, derivations are ruled out. The probability of this transition itself can be defined as a ratio of sums, \(\frac{\sum_{\text{after}}}{\sum_{\text{before}}}\). The log of the reciprocal of this ratio is the surprisal of the next word. The greater the probability mass ruled out, the higher the surprisal.

\(^6\)The distributions illustrated in Figure 3 have not been renormalized; in this sense the picture is drawn from the point of view of the grammar which assigns probabilities to full derivations.
5.2 Empirical support for surprisal

Surprisal has seen broad success across many different methodologies. In eye-tracking, several different parsing mechanisms and grammar types converge on the idea that people read more slowly on words whose surprisal value is higher [Boston et al., 2008, Demberg and Keller, 2008, Rauzy and Blache, 2012, inter alia]. With event related potentials, similarly, surprisal is positively related to the amplitude of the N400 component [Frank et al., 2015]. And in functional MRI, surprisal from phrase-structured language models predicts the timecourse of activation in several different brain areas including anterior temporal lobe [Hale et al., 2015, Brennan et al., in revision, Henderson et al., under review]. Timecourses from several other brain areas appear to correlate well with surprisal values from n-gram models [Willems et al., 2015].

Surprisal is successful empirically because it accounts naturally for frequency effects. Such effects have long been recognized in behavioral studies. To take just one example, Thibadeau et al. [1982] found unigram probability to be a useful predictor of eye-fixation duration. Surprisal generalizes this idea from single words to larger domains, e.g. where \( n > 1 \). It extends the idea that “rare-implies-more-difficult” to syntactic phrases. However, there’s more than one way to do it! In making this extension, several design decisions present themselves. Many of these reflect classic issues in cognitive science for which the complexity metric itself offers no easy answers. The following subsections take up two such issues as they relate to surprisal.

5.3 Crosscutting design decisions

5.3.1 Lexicalization

Linguistics has wrestled since the 1970s with lexicalism, the idea that grammatical properties should be sensitive to the idiosyncrasies of words themselves. The decision to be lexicalist or not carries over directly into information-theoretical complexity metrics. For instance, the phrase-structured language model of section 4.2 is partially lexical. This means that some probabilities are contributed by preterminal rules having to do with real words e.g. \( N \rightarrow beer \) or \( V \rightarrow loves \) while others come from phrasal rules like \( NP \rightarrow Adj \ N \) or \( PP \rightarrow P \ NP \) that refer only to syntactic categories. If the research question focuses on syntactic processing in its own right, then one can fit the grammar or parser to part of speech tags, rather than real words. This suppresses the lexical contribution.

Rather that suppressing lexicalism, the other possibility is to embrace it. One can lexicalize more and more rules, including those classically viewed as independent of particular words [see the example in Hale, 2014, chapter 2]. Such an approach holds out hope of accounting for detailed lexical effects in human sentence processing [as suggested by Ford et al., 1982, MacDonald, 1994]. However, it does so by entangling the notion of predictability with that of rare words. These two factors seem to be psychologically separable and might be better understood in isolation from each other [Staub, 2015]. Another downside is the greater number of parameters in lexicalized grammars. Allowing for more interactions between phrases and words makes them more difficult to estimate in practice.

5.3.2 Parallelism

Another key design decision revolves around approximations of the prefix probability. In the Hale [2001] formulation, all structural alternatives affect the metric. One might interpret this in terms of a fully parallel parsing mechanism that considers all possible analysis paths. Under this regime an analysis remains in play until it is definitively ruled-out by some observed word. This is exactly what happens in garden path sentences, where the “pain” of committing to the globally-correct
interpretation is deferred until the disambiguation point. In this way, surprisal can emulate theories of attachment preferences, such as Frazier and Fodor’s Garden Path Theory. An alternative is to impose some parallelism limit. Parsing accuracy comparisons suggest that even a restriction to 3 or 4 syntactic analyses leads to good performance on the Wall Street Journal [Brants and Crocker, 2000]. In light of this, Boston, Hale, Vasishth, and Kliegl [2011] parametrically varied the amount of parallelism available to a particular parser. The eyetracking predictions derived in this manner got better and better as the parser considered more alternative pathways. This convergence toward an ideal, where memory is unlimited, suggests that the metric itself is on the right track.

5.4 Empirical challenge: relative clauses

On the other hand, there is data suggesting that surprisal’s simple equation between probability and difficulty may actually be too simple. The empirical challenge comes from relative clauses (RCs). In a relative clause there is a missing element, marked \( t \) in example (1) below.

(1) a. The reporter that \( t \) attacked the senator admitted the error
b. The reporter that the senator attacked \( t \) admitted the error

In transformational grammar, one would say that the missing element has been “relativized” leaving a “trace.” In a particular language, relativization applies to a delimited subset of grammatical relations: subject, direct object, indirect object et cetera. These subsets, to which relativization may apply, are organized into a scale such that processing difficulty varies inversely with typological ubiquity in the world’s languages [Keenan and Hawkins, 1987]. Within psycholinguistics, attention has focused on the first two points of this scale, subject- and object- extracted RCs. These are shown in example (1) using stimuli from King and Just [1991].

Object-extracted RCs as in (1-b) are rare in natural language corpora. Probabilistic grammars fitted to them end up including a low-probability rule. In virtue of such a rule, surprisal correctly predicts that (1-b) should be the more difficult of the two constructions. But as an incremental complexity metric it is less accurate: the vanilla version of surprisal would predict effort at the point where the low-probability rule becomes obligatory. This happens at the left-edge of the relative clause, not at the embedded verb \( attacked \). This is where Grodner and Gibson [2005] report a reading time slowdown. While Staub [2010] later found effects at this leftmost position, the gen-

---

7 Frazier and Clifton [1996, chapter 1] provide a concise summary of Garden Path Theory. For the standard formalization due to Pereira and Shieber, see Hale [2014, chapter 4].

8 Bresnan and Kaplan [1982] exhort cognitive scientists to “discover ways of showing that the actual behavior of real native speakers converges on the ideal behavior predicted by our grammatical theory, as interfering performance factors are reduced” [page xxiii]. Kaplan later explained this further, saying:

The basic idea is that you can evaluate theories of grammar-based processing as to whether their behavior corresponds to the behavior of an ideal native speaker in the limit as the amount of available processing resources goes to infinity. Of course, the behavior of an ideal native speaker, one who knows his language perfectly and is not affected by restrictions of memory or processing time, lapses of attention, and so forth, is difficult to observe. But as psycholinguistic methods and technologies improve, we can imagine doing experiments in which we somehow vary the cognitive resources of real speakers and hearers, by removing distractions, giving them scratch-pad memories, etc. We can then take the limiting, asymptotic behavior of real speakers as approximations to the behavior of the ideal. A grammar-based processing model which, when given more and more computational resources, more and more accurately simulates the behavior of the ideal has the ‘ideal-convergent’ property.

[Kaplan, 1995, page 344]
eral feeling during the early 2000s was that surprisal had failed to derive an important part of the data pattern. Expressing this general feeling, Levy [2008a, page 1166] refers to “mixed results.”

One way of interpreting these mixed results is to hypothesize that surprisal has a major effect on word-by-word processing difficulty, but that truly non-local (i.e., long-distance) syntactic dependencies such as relativization and WH-question formation are handled fundamentally differently [...]

The suggestion is to back off to a more complex, two-factor model where surprisal’s role is somewhat curtailed.9

It is of course possible that some other grammar type or parsing mechanism would yield better surprisal predictions, by specifying a different order in which the relevant probabilities take effect. But investigation along this line faltered under the assumption that the embedded verb is the primary locus of processing difficulty. This mismatch between theory and data motivated the search for another complexity metric, one that would provide a one-factor explanation of the difficulty profile in relative clauses.

6 Entropy reduction

6.1 Definition

Entropy reduction is that metric. Unlike surprisal’s logarithmic transformation of probability, it instead formalizes the information value of a word using the notion of entropy. This quantity is the centerpiece of information theory. Shannon [1948] suggests its name by analogy to thermodynamics.

The quantity which uniquely meets the natural requirements that one sets up for “information” turns out to be exactly that which is known in thermodynamics as entropy.

The entropy of a random variable \( X \) is defined below in equation 3.

\[
H(X) = - \sum_{x \in X} P(x) \log_2 P(x) \tag{3}
\]

The entropy \( H \) quantifies uncertainty in \( X \)’s probability distribution. For instance, a 100-sided die has greater uncertainty than a 6-sided die. When outcomes are equiprobable, entropy is at a maximum. When alternatives are unequally weighted, it is easier to guess the outcome; we become less uncertain. The core intuition of the Entropy Reduction complexity metric is that this sort of ‘information gain’ should index observable human comprehension difficulty.

Entropy Reduction was inspired by, and can be viewed as a generalization of Den and Inoue’s [1997] Verb Predictability Hypothesis. Whereas Den and Inoue were concerned with the size of the garden path effect at a sentence-final verb, Entropy Reduction applies to all positions and all categories. Taking \( X \) again to be derivations on a probabilistic grammar, we ask: by how much does knowledge of the initial substring \( Y = y \) reduce uncertainty? The answer is the information value \( I \)

\[
I(X; y) = H(X) - H(X|y) \tag{4}
\]

In Equation 4 the quantity \( H(X|y) \) is simply the entropy of the subset of derivations sharing an initial substring. Analogous to the way prefix probabilities were divided in section 5.1, subtractions

---

9Demberg and Keller [2009] go on to propose just this sort of two-factor model.
of these conditional entropies $H(X|y)$ define the informational contributions of particular words. If the conditional entropy goes down from one word to the next, then grammatical uncertainty has been reduced. The interpretation is that the comprehender has done information-processing work. On the other hand, it is also possible for entropy to go up. This happens when the next word opens up possibilities that are more uncertain than on average. In such cases, the comprehender has done no work: no progress has been made towards the goal of a unique reading.\footnote{An approximation of entropy reduction restricts consideration to just uncertainty about the next word, rather than the whole derivation. Here again, a theorist must decide how to structure the left context: either by phrase, as in Roark et al. [2009] or by word, as in Willems et al., 2015.}

Figure 4 repeats Figure 3 in order to graphically emphasize the point that surprisal and entropy reduction yield different information values, even from the same grammatical probability distributions. The two metrics summarize the transition between distributions in different ways. Appendix A provides a worked example showing how contrasting predictions follow from the same grammar. These specific cases exemplify the general point, developed in Blachman [1968], that entropy reduction and surprisal are different conceptions of information value.

Figure 4: The same “before” and “after” distributions as in Figure 3 yield different information values. Entropy reduction is the downward change, if any, between $H$(before) and $H$(after), whereas surprisal is the log-ratio of sums of these distributions.

6.2 Empirical support

Entropy reduction derives the comprehension difficulty profile across a wide range of constructions, including relative clauses [Hale, 2003, 2006, Yun et al., 2015, Chen and Hale, 2015]. It has been applied with naturalistic text stimuli [Wu et al., 2010, Frank, 2013] as well as with controlled experimental materials [Linzen and Jaeger, In press]. Beim Graben and colleagues [2000, 2008] show how ERP components such as the N400 and P600 can be understood as entropy reductions in an underlying dynamical system.

Using techniques due to Bar-Hillel et al. [1964] and Nederhof and Satta [2008], entropy reduction can be computed for a wide array of formal grammars. Freely-available software [Chen et al., 2014]
facilitates this computation, even with expressive grammars that directly define syntactic movement. Using the “promotion analysis” of relativization, for instance, Yun et al. [2015] model the difficulty profile of Chinese, Japanese and Korean relative clauses. In these languages, relativization dependencies are arranged in such a way that their processing asymmetry cannot be explained by simple memory-based approaches. However, a grammatical approach — where processing effort is observable at the point where uncertainty is reduced — does work. The same syntactic analysis thus yields an account of the processing difficulty profile in both prenominal as well as postnominal RCs [Hale, 2006].

On a conceptual level, entropy reduction takes one step away from the externalism of surprisal by focusing on an internal aspect of parser states, their uncertainty. From this point of view, the distribution over alternative analyses matters in a way that it does not for surprisal. This research program is naturally extended by adding more information about an agent’s goals, such that these internal probabilities align with expected payoffs, becoming “utilities” [Hale, 2011, Lewis et al., 2013, Calvillo and Crocker, 2015].

7 Mechanisms

It is foundational, in cognitive science, to differentiate between levels of explanation. Marr [1982] distinguishes between an higher “computational” level and a lower “algorithmic” level. These two levels are simultaneous co-descriptions of the same organism, at greater and lesser degrees of abstraction. Theories that define what computation the organism is actually doing are stated at the higher level, while theories of how that computation is effected occupy the lower level.

Complexity metrics like Entropy Reduction and Surprisal, then are computational-level theories: they specify what the difficulty level of parsing a word should be in terms of structures defined in a language model. In order to offer a psychological process model, one must characterize a mechanism that operates in accordance with these metrics — a compatible co-description.

Hale [2014, chapter 8] advances a mechanistic interpretation of surprisal, viewing it as a consequence of the Chunking Theory of Learning (CTL) [Rosenbloom and Newell, 1987]. The CTL supposes that cognitive operators can fuse together in the course of practice. In this way, what used to require multiple steps is now accomplished by a macro-operator that applies all at once, in one step. Analyzing the CTL, Rosenbloom and Newell show how it necessarily derives power-law practice curves, in environments where the probability of a “pattern” decreases as the pattern size increases. This is certainly true in natural language. They remark (page 227) upon the ease with which a power law can be mimicked by a purely logarithmic function. Of course, this is exactly what surprisal is (Equation 1).

Support for this reduction of surprisal to the CTL comes from analyses of eyetracking data in Hale [2014]. These analyses considered triplets of parser actions, aligned in time to particular words in English and French newspaper text. An independent estimate of the degree to which these triplets “cohere”, i.e. were likely to be chunked by a chunking mechanism, was a positive predictor of eye-fixation duration. The interpretation is that a person reads familiar sentence structures faster because the cognitive operations that build those structures have been folded into macro-
operators. From this perspective, construction grammar would be a correct description of the chunked representations at the same time as generative grammar remains a correct description of the un-chunked representations. Further work along this line might consider alternative training regimes that experimentally induce chunking.

8 Past and Future

The revival of information-theoretical complexity metrics is a revival of ideas that fascinated information-processing psychologists like Hick [1952], Attnave [1959] and Garner [1962] as well as pre-generative linguists like Charles F. Hockett [1953]. In the 1950s, information theory seemed to offer limitless new vistas for understanding human language. But as Luce [2003] details, it fell out of favor. Among other factors, Luce places the blame on a lack of structure in the models that were at that time under consideration:

The elements of choice in information theory are absolutely neutral and lack any internal structure. That is fine for a communication engineer ....[but] by and large, however, the stimuli of psychological experiments are to some degree structured, and so, in a fundamental way, they are not in any sense interchangeable.

In the 21st century, we now know how to build structured probabilistic models. Grammars are just one example of a probabilistic model that is at the same time sensitive to the statistics of the environment and also able to generalize in a categorical, rule-governed way. Now, unlike in the 1950s, we can assign both an information-theoretic interpretation (e.g. the choice of this symbol is worth x bits) as well as a detailed linguistic interpretation (this rule introduces a relative clause) to the same mathematical entity. By combining contributions from disciplines that are traditionally separate, we advance cognitive science.

Acknowledgement

This research was supported by NSF CAREER award 0741666.
A Contrasting predictions

The same grammar can lead to different predictions via entropy reduction and surprisal, respectively. This section develops an example where that happens.\textsuperscript{14}

\[
\begin{align*}
0.98 & \quad S \rightarrow A \; X \quad X \text{ is radically more probable than } Y \\
0.01 & \quad S \rightarrow B \; Z \quad Z \text{ is a fairly uncertain category} \\
0.01 & \quad S \rightarrow C \; Y \quad \text{choice between } X \text{ and } Y \text{ is cued by first symbol, } a \text{ vs } c \\
1.0 & \quad A \rightarrow a \\
1.0 & \quad B \rightarrow b \\
1.0 & \quad C \rightarrow c \\
1.0 & \quad X \rightarrow f \quad \text{neither } X \text{ nor } Y \text{ is entropic} \\
1.0 & \quad Y \rightarrow g \\
0.25 & \quad Z \rightarrow c \\
0.25 & \quad Z \rightarrow d \\
0.25 & \quad Z \rightarrow e \\
0.25 & \quad Z \rightarrow f
\end{align*}
\]

Consider the surprisal value of strings starting with $c$ and $b$. Their surprisals are exactly the same, 6.64 bits. Intuitively, both of them eliminate the highly-probable $S\rightarrow A \; X$ rule in favor of one or another lower-probability rule.

The entropy reductions associated with these symbols are different: 0.1814 versus none at all. If the first symbol is $c$, then the rule $S\rightarrow C \; Y$ will be required. Since (by construction) there is no uncertainty at all about $Y$‘s derivation, this means that all of the entropy associated with $S$ has been reduced.

If instead the first symbol is $b$, then the rule $S\rightarrow B \; Z$ will be required. There is now and “obligation” to work out $Z$‘s derivation. Since $Z$‘s derivation is maximally uncertain, entropy goes up: no work is done on this particular symbol. On the following symbol, where $Z$‘s derivation is revealed, all of this entropy is reduced.

\[
\begin{array}{c|c|c}
\text{symbol} & \text{prefix prob} & \text{c} \\
\hline
\text{prefix prob} & 1 & 0.01 \\
\text{surprisal} & - & 6.64 \\
\text{entropy} & 0.1814 & 0 \\
\text{entropy reduction} & - & 0.1814 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
\text{symbol} & \text{prefix prob} & \text{b} \\
\hline
\text{prefix prob} & 1 & 0.01 \\
\text{surprisal} & - & 6.64 \\
\text{entropy} & 0.1814 & 2 \\
\text{entropy reduction} & - & \text{none} \\
\end{array}
\]

\textsuperscript{14}Final version will have thank-you here in connection with the example
References


